11-27-2015 12:00 AM

Efficient Scheduling Algorithms for Wireless Resource Allocation and Virtualization in Wireless Networks

Mohamad Kalil
The University of Western Ontario

Supervisor
Abdallah Shami
The University of Western Ontario

Graduate Program in Electrical and Computer Engineering
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Efficient Scheduling Algorithms for Wireless Resource Allocation and Virtualization in Wireless Networks

(Thesis format: Monograph)

by

Mohamad Kalil

Graduate Program
in
Electrical and Computer Engineering

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

Abstract

The continuing growth in demand for better mobile broadband experiences has motivated rapid development of radio-access technologies to support high data rates and improve quality of service (QoS) and quality of experience (QoE) for mobile users. However, the modern radio-access technologies pose new challenges to mobile network operators (MNO) and wireless device designers such as reducing the total cost of ownership while supporting high data throughput per user, and extending battery life-per-charge of the mobile devices. In this thesis, a variety of optimization techniques aimed at providing innovative solutions for such challenges are explored.

The thesis is divided into two parts. In the first part, the challenge of extending battery life-per-charge is addressed. Optimal and suboptimal power-efficient schedulers that minimize the total transmit power and meet the QoS requirements of the users are presented. The second outlines the benefits and challenges of deploying wireless resource virtualization (WRV) concept as a promising solution for satisfying the growing demand for mobile data and reducing capital and operational costs. First, a WRV framework is proposed for single cell zone that is able to centralize and share the spectrum resources between multiple MNOs. Consequently, several WRV frameworks are proposed, which virtualize the spectrum resource of the entire network for cloud radio access network (C-RAN)- one of the front runners for the next generation network architecture.

The main contributions of this thesis are in designing optimal and suboptimal solutions for the aforementioned challenges. In most cases, the optimal solutions suffer from high complexity, and therefore low-complexity suboptimal solutions are provided for practical systems. The optimal solutions are used as benchmarks for evaluating the suboptimal solutions. The results prove that the proposed solutions effectively contribute in addressing the challenges caused by the demand for high data rates and power transmission in mobile networks.

Keywords: Power Minimization, QoS, LTE, Packet Scheduling, Uplink, Virtualization, C-RAN, Resource Sharing

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Abdallah Shami, for the immeasurable amount of support, guidance, and personal and professional advices he has provided me while pursuing this degree. I could not imagine having a better supervisor for my PhD study. I appreciate all his contributions of time, ideas and support to make my PhD experience productive, successful and enjoyable. Thank you for motivating me along the way and for being the go-to person whenever I faced problems or felt down. I knew I could always turn to him when my depleted motivation needed recharging.

My sincere gratitude goes to my co-supervisor Dr. Arafat Al-Dweik for his continuous help and support in all stages of my PhD study. Without his involvement, this work will not have been possible. His encouragement, kindness, brilliant comments and insightful suggestions, and his prompt responses to my questions will always be remembered.

I would like to thank my examiners: Dr. Serguei Primak, Dr. Lian Zhao, Dr. Aleksander Essex, and Dr. Michael Bauer for taking the time to review and examine my thesis, and for their insightful comments and suggestions.
I would also like to thank my Master's thesis supervisor, Dr. Mohammad Banat, for helping me to begin my research journey. I would like to thank Dr. Ibrahim M. Ghareeb and Dr. Redha Radaydeh, who taught me the first courses in wireless communication and introduced me to the area.

Special thanks go to my colleagues and friends at Western university who made my PhD years a joyful and memorable experience. For most, a BIG thanks to Mohamed Abu Sharkh who experienced all of the ups and downs of my research, and to Fuad Shamieh who was always there for me and ready to help. Also, I would like to extend my thanks to my friends: Oscar Filio, Karim Hammad, Aidin Reyhani, Dan Wallace, , Manar Jammal, Hassan Hawilo, Bradley de Vlugt, Elena Uchiteleva, Abdallah Moubayed, Khaled Alhazmi, Emad Aqeeli, Mohamed Hussein, Maysam Mirahmadi, Khalim Meerja, Siamack Ghadimi, M. Ajmal Khan, Eric Southern, Peng Hao, Mohamed Youssef, Abdulfattah Noorwali, Mohammad Noor Injadat, Fadi Salo, Anas Saci, Marco Luccini, Anas Ibrahim, Jay Nadeau, and Kevin Mi.

To my friends scattered around the world, thank you for your well-wishes, phone calls, e-mails, and being there whenever I needed a friend. For most, I would like to thank Yazan Al-Badarneh, Khair Al Shamaileh, Majdi Ababneh, Alaeddin Bani Milhim, Ali

Bani Amer, Yousef Mashaala, Bashar Alwadyan, Malek Khudirat, Qutaiba Al-Hazaime, Mohamad Abu Hani, Hasan Thiabat, and Ahmad Al-Sharoa.

A warm thanks goes to Madeleine and Carl, whom I lived with for almost 3 years, for being warm, kind, and a second family for me.

And most importantly, I am deeply and forever indebted to my family. You are the greatest source of love, encouragement, and inspiration. In particular, this project is dedicated to them with my sincerest thanks and appreciation to my father, God bless his soul, who passed away last may, and to my mother for their love, support and encouragement throughout my entire life, and for all of the sacrifices they have made on my behalf. Without them I would have been totally lost and never took a step ahead.

My sincere gratitude goes to my sisters Maysa, Mayada, Maram, and Marwa, and my brothers, Motaz and Mones, as well as to my grandmothers, grandfathers, father-in-law, mother-in-law, brothers-in-law, sister-in-law, aunts, uncles, and cousins for their love and support, and for being proud of me.

Last but not least, a heartfelt thanks to my wife Tasneem for her continued support, love, and for being my best friend and great companion who helped me get through the unexpected troubles of research in the most positive way. Words cannot describe how lucky I am to have you in my life.

To the memory of my father, and to my mother and my loving wife Tasneem

Table of Contents

List of Tables

List of Figures

Acronyms

3GPP	*3rd Generation Partnership Project*
ACG	*Average Channel Gain*
APF	*Access Proportional Fairness*
ARP	*Allocation and Retention Priority*
AWGN	*Additive White Gaussian Noise*
BIP	*Binary Integer Programming*
BLER	*Block Error Rate*
BPU	*Baseband Processing Unit*
BPWR	*Bit per Watt Ratio*
BS	*Base Stations*
BSR	*Buffer Status Reporting*
BWC	*Bit per Watt Controlling*
CAPEX	*Capital Expenditure*
CoMP	*Coordinated Multipoint*
CPU	*Central Processing Unit*
C-RAN	*Cloud Radio Access Network*
DP	*Dynamic Programming*
DU	*Data Unit*
eNB	*Evolved Node-B*
FFR	*Fractional Frequency Reuse*
FR	*Frequency-Reuse*
GBR	*Guaranteed Bit Rate*
GPF	*Generalized Proportional Fair*
GWCN	*Gateway Core Network*
HetNet	*Heterogeneous Network*
HoL	*Head-of-Line*

ICI	*Intercell Interference*
ICIC	*Intercell Interference Coordination*
LSM	*Least Satisfied Mobile Network Operator*
LTE	*Long Term Evolution*
MAC	*Media Access Control*
MATP	*Maximum Allowable Transmit Power*
MCS	*Modulation And Coding Scheme*
MDP	*Markov Decision Process*
MIMO	*Multiple-Input Multiple-Output*
M-LWDF	*Modified-Largest Weighted Delay First*
MME	*Mobility Management Entity*
MNO	*Mobile Network Operator*
MOCN	*Multi-Operator Core Network*
MT	*Maximum Throughput*
Multi-RAT	*Multiple Radio Access Technology*
MWIS	*Maximum Weighted Independent Set*
NGBR	*Non Guaranteed Bit Rate*
NGN	*Next Generation Network*
OFDMA	*Orthogonal Frequency Division Multiple Access*
OPEX	*Operating Expenditure*
PAPR	*Lower Peak-To-Average Power Ratio*
PDB	*Packet Delay Budget*
PDF	*Probability Density Function*
PDU	*MAC Protocol Data Units*
PF	*Proportional Fair*
PHY	*Physical Layer*
QCI	*Quality of Service Class Identifier*
QoS	*Quality of Service*
RAP	*Resource Block Access Probabilities*
RB	*Resource Block*
RLC	*Radio Link Control*
RRH	*Remote Radio Head*

SC-FDMA	*Single Carrier Frequency Division Multiple Access*
SINR	*Signal-to-Interference Plus Noise Ratio*
SLA	*Service-Level Agreement*
SNR	*Signal to Noise Ratio*
SON	*Self-Organizing Network*
SP	*Service Providers*
SS	*Static Sharing*
TTI	*Transmission Time Interval*
UE	*User Equipment*
vMNO	*Virtual Mobile Network Operator*
WRV	*Wireless Resource Virtualization*

Chapter 1
Introduction

The demand for mobile data is growing exponentially and mobile network operators (MNOs) are struggling more than ever to increase their capacity profitably. Mobile communication is becoming an essential need for individuals and businesses. The number of new mobile-connected devices added in 2014 reached almost half a billion, meaning the total number of mobile-connected devices exceeds the world's population [4]. Mobile Internet-access is also visibly becoming affordable and available for a larger segment of the population. This leads to Internet usage levels we thought were once unreachable. A look at the Youtube statistics shows that more than 4 billion hours of video are watched on YouTube every month. Pressures on the infrastructure, service levels and performance have never been that high. The demand for mobile data does not stop there. Mobile data traffic forecasts estimate a 10-fold increase in global mobile data traffic between 2014 and 2019. This unprecedented penetration is accompanied by a major increase in mobile network connection speeds. For example, the average mobile network downstream speed increased 20 percent in 2014 [4].

At the same time, the high demand for mobile data has a significant impact on the power consumed by mobile devices because their transmission power substantially increases with the increase in uplink data rates. Furthermore, the increasing popularity of using online storage services, social media, and social networking applications on Smartphones boosts the uplink traffic volumes. Thirty billion pieces of content are shared on Facebook every month. Users do not only view downloaded content, they also contribute to creating content by uploading photos and videos. Uplink usage is gaining more significance to the extent that, according to AT&T Labs Research, the uplink and the downlink traffic volumes become similar in the case of large events like the Super Bowl [5]. All of this demand is countered by an annual gain in battery capacities of just 4 percent [6]. Improvements in the battery capacity of mobile devices lag far behind the surging demand for

transmission power, which makes power-efficient communication in uplink transmission an essential requirement for next-generation mobile networks.

Due to the large investments in network resources needed to support the surge in mobile data traffic, MNOs profits have not been growing at rate that the traffic volume level would indicate. In their continuous endeavors to tackle this issue, MNOs have become highly interested in cost-effective solutions in order to satisfy the high demand for mobile data. To maximize the average revenue per user, MNOs have to efficiently utilize the limited and highly expensive spectrum resources. A surprising fact here arises. Recent spectrum utilization measurements have shown that the bandwidth licensed to MNOs is mostly underutilized. According to a recent study by Nokia [7], only 20 percent of a radio access network's full capacity is used at any given time with 80 percent being idle and waiting for peak hour demand. This motivates innovative solutions to efficiently use the limited spectrum resources and maximize profits.

1.1 Thesis Outline and Contributions

The rest of this thesis is divided into five chapters. Chapter 2 presents two QoS-aware power-efficient schedulers for mixed streaming services in long term evolution (LTE) uplink systems. Chapter 3 focuses on finding a global optimal scheduler that minimizes the total transmission power while satisfying particular delay requirements in the LTE uplink. Moreover, two low-complexity heuristic schedulers are proposed to solve the optimal formulation with comparable power consumption and comparable performance. It is worth mentioning that chapter 2 optimizes the transmission power every transmission time interval (TTI). Nevertheless, chapter 3 finds the global optimal solution by optimizing the transmit power over multiple TTIs. In chapter 4, the concept of wireless resource virtualization (WRV) is introduced. Furthermore, an efficient low-complexity WRV scheduler that maintains access proportional fairness between users and MNOs at a single base station is derived. Chapter 5 considers network-wide virtualization. Optimal and suboptimal solutions for virtualizing cloud radio access networks (C-RANs) wireless resources are provided. Finally, Chapter 6 concludes the thesis and proposes future directions for extending the research.

1.2 Contributions of the Thesis

The major contributions of the thesis are summarized as follows.

1.2.1 Contributions of Chapter 2

1. An optimal formulation of the resource allocation problem in the LTE uplink is presented. In contrast to previous work, the formulation considers the maximum transmission power threshold of the users, which may produce infeasible solutions. In addition, the challenge of solution infeasibility is addressed by amending the objective function with additional terms.
2. The proposed schedulers are designed and evaluated for a realistic LTE framework, where each user requests many bearers with different QoS requirements.
3. A low complexity iterative resource allocation algorithm is derived to solve the optimal formulation with comparable power consumption and comparable performance. Moreover, the computational complexity of the proposed schedulers is analyzed. The algorithm is $O(MK)$, where M is the numbers of users and K is the number of RBs.

1.2.2 Contributions of Chapter 3

1. The global optimal scheduler is derived for the LTE uplink. The scheduler minimizes the total transmit power of all users while satisfying delay requirements. Unlike Chapter 2, where the system is optimized for one time slot, the optimization framework presented in this chapter takes into consideration several time slots, and finds the minimum transmit power that satisfies the delay requirements of the users. This solution can be used as a benchmark for comparing different schedulers used for the LTE uplink.
2. The scheduling problem is formulated as a dynamic programming (DP) problem, and the scheduler considers the dynamic nature of the traffic load, maximum transmit power threshold, contiguous allocation, and the time-varying fading channel.

3. To reduce complexity, we propose two power-efficient heuristic schedulers to solve the scheduling problem. The algorithms are $O(K \times M \times log(M))$. The performance of the heuristics is evaluated and compared to the optimal scheduler and to other existing schedulers.

1.2.3 Contributions of Chapter 4

1. The main benefits of WRV and an overview of existing WRV techniques are provided and classified.
2. An efficient low-complexity WRV scheme is provided. The scheme aims at maximizing the throughput while maintaining access proportional fairness (APF) among users and MNOs.
3. An iterative offline algorithm is developed to compute the optimal weights that the scheme uses to maintain the APF. The algorithm has low complexity, converges after a small number of iterations, and it is scalable to large-scale scenarios.

1.2.4 Contributions of Chapter 5

1. An optimal scheme that enables spectrum sharing between multiple MNOs and RRHs is proposed and formulated. The scheme eliminates intercell interference (ICI) between RRHs and considers fair distribution of spectrum resources between RRHs based on their traffic loads. Moreover, it provides a high level of isolation between MNOs, allows different MNOs to apply different customizable resource scheduling policies, and offers efficient resource utilization across the network.
2. A suboptimal scheme that solves the optimization problem with lower complexity is derived. The suboptimal scheme is obtained by dividing the wireless resource allocation problem into sub-problems. The objective of each sub-problem is to allocate one RB to a set of non-interfering RRHs. The allocation per single RB is formulated as a maximum weighted independent set problem, which is solved using binary integer programming (BIP) solvers.
3. To further reduce the complexity, a low-complexity heuristic algorithm that solves the BIP problem is proposed. The proposed heuristic algorithm is greedy and finds

the set of non-interfering RRHs for each RB iteratively. The time complexity of the heuristic algorithm is considerably lower than the BIP scheme while its throughput and delay performance are comparable.

Chapter 2
QoS-Aware Power-Efficient Scheduler for LTE Uplink

2.1 Introduction

The continuous introduction of mobile applications and services leads to a significant increase of data usage over mobile devices. To accommodate the drastic growth of mobile data traffic and improve the system capacity, long-term evolution (LTE) technology has been developed by the 3rd generation partnership project (3GPP). LTE provides superior speed, low latency and better quality of service (QoS) for mobile networks. The target peak cell aggregated downlink throughput within a 20 MHz spectrum can reach 300 Mbit/s in downlink, and 75 Mbit/s in uplink by applying multiple-input multiple-output (MIMO) configurations [9]. However, the high speed data links offered by LTE systems increase the power consumption of the user equipments (UEs) [10]. In uplink mobile communication systems, the power source of the UE is limited to a battery. Nevertheless, the improvement of battery technologies is much slower than the steadily rising demand for transmission power by UEs. Consequently, the battery life per charge is currently one of the main factors that dominate the reliability of mobile devices.

To enhance the power efficiency of UEs, LTE uses single carrier frequency division multiple access (SC-FDMA) in the uplink, while orthogonal frequency division multiple access (OFDMA) is used for the downlink. SC-FDMA has a lower peak-to-average power ratio (PAPR) when compared to OFDMA. The low PAPR advantage allows the power

---

A version of this chapter has been published in [8].

amplifier at the transmitter to operate close to the saturation point which improves its efficiency. However, the lower PAPR feature of SC-FDMA requires contiguous allocation of the resource blocks (RBs) [9].

In this chapter, power-efficient schedulers for LTE uplink systems are proposed. The proposed schedulers are able to minimize the total instantaneous transmission power from all users while maintaining the LTE uplink physical layer constraints and the QoS requirements. It is noteworthy that minimizing the total instantaneous transmission power can lead to minimizing the average transmission power specially in burst transmission.

The rest of the chapter is organized as follows. Section 2.2 presents the related work. Section 2.3 presents the system model. The system constraints and the objective of the work are presented in Section 2.4. The optimal and iterative formulations of the proposed scheduler are described in Sections 2.5 and 2.6, respectively. In Section 2.7 the PF scheduler is discussed. Intra-user scheduling is described in Section 2.8. Simulation results are presented in Section 2.9, and Section 2.10 concludes the chapter.

2.2 Related Work

The resource allocation problem for OFDMA systems has been widely investigated in the literature [11, 12]. As each RB cannot be assigned to more than one user, the resource allocation is a combinatorial optimization problem [11], which cannot be solved in polynomial time. Many studies have solved the allocation problem by using the convex relaxation method [11, 12]. The relaxation replaces the binary variables by continuous variables in the interval $[0, 1]$, then Lagrange multipliers are derived to solve the resource allocation. However, due to the contiguity constraint in SC-FDMA, the resource allocation methods that are applied to OFDMA are not directly applicable to the LTE uplink [13]. More details about the convex relaxation method are discussed in Section 2.4.

Most recent research efforts on LTE uplink have addressed the maximization of the total utility of the system [14, 15, 16, 17]. The utility function may depend on users' throughputs, maximum permitted delays, or fairness between users. A sum-rate maximization problem is investigated in [14]. The allocation problem is formulated as a set packing

problem which is NP-hard. Consequently, low-complexity algorithms based on message-passing paradigm are proposed to solve the allocation problem in polynomial time. Another sum-rate maximization scheduler that considers multiuser scheduling with transmit antenna selection is investigated in [15]. Based on local ratio test approach, suboptimal polynomial-time algorithms are proposed to solve the allocation problem. A scheduling approach based on a genetic algorithm is presented in [16] to solve the sum-rate maximization problem in LTE uplink systems. Nevertheless, sum-utility maximization problems usually lead to transmitting data using the maximum allowable transmission power, which often lowers the transmitter power efficiency [18].

A general packet scheduling scheme for LTE uplink is considered in [17]. The problem is proven to be MAX SNP-hard. Consequently, two approximation algorithms for the scheduling problem are proposed to reduce complexity. The algorithms are evaluated for a specific scenario that incorporates the queue length and channel quality information. Lee *et al.* [19] investigated the proportional fair (PF) scheduler for the LTE uplink. The scheduling problem is proven to be an NP-hard because of the contiguity constraint. Heuristic algorithms were proposed and compared in terms of system throughput and fairness.

Few articles in the literature have considered power-efficient transmission in the LTE uplink. For example, Dechene *et al.* [20, 21] considered power-efficient resource allocation subjected to rate and synchronous hybrid automatic repeat request constraint. At every transmission time interval (TTI), each user transmits a fixed data rate. The solution feasibility was guaranteed because maximum transmit power is not considered. However, practical LTE uplink systems limit the maximum transmission power to a threshold value, and the data arrives randomly to the users' buffers with different QoS requirements. Another power-efficient scheduler that considers the buffers' queue state information is presented in [22]. The problem was formulated as a constrained Markov decision process (MDP), offline solutions were then derived. However, the maximum transmission power threshold is not considered. Besides, the solution complexity is high due to the large size of the search space. Furthermore, the offline solution is applicable only to a particular scenario, meaning that different solutions must be derived for different scenarios. For example, adding a user to the system leads to a different solution.

2.3 System Model

This study considers an LTE uplink multiuser system in a single cell, where K UEs communicate with an evolved node-B (eNB). It is assumed that each user has a maximum of four bearers, or logical channels, associated with different QoS requirements, an assumption which is justified in Section 2.3.1. The overall cell bandwidth is divided equally into M RBs, each of which contains 12 adjacent subcarriers. The bandwidth of an RB is 180 kHz. To facilitate the readability, Table 2.1 summarizes the notations frequently used throughout the chapter.

The contiguity constraint which is required by SC-FDMA can be modeled by constructing a binary matrix $\mathbf{R}$ as follows. Each column in $\mathbf{R}$ represents a potential contiguous allocation, while each row represents an RB. The column index c in the matrix $\mathbf{R}$ with the size $M \times 1$ is denoted by $\mathbf{r}_c$. Therefore, $\mathbf{R}$ can be expressed as

$$\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_C]. \tag{2.1}$$

where C is the number of columns in $\mathbf{R}$, which can be calculated for a system that has M RBs as [20]

$$C = \frac{1}{2}M(M+1). \tag{2.2}$$

For example, a system with 3 RBs has an $\mathbf{R}$ that is equal to

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}. \tag{2.3}$$

To better understand the meaning of allocating one column of $\mathbf{R}$ to a user, consider that the column $\mathbf{r}_1$ is allocated to user k. The first element in $\mathbf{r}_1$ is one, which indicates that RB number one is allocated to user k. The second and third elements in $\mathbf{r}_1$ are zeros, which indicates that RB numbers two and three are not allocated to user k. As can be seen in (2.3), $\mathbf{R}$ contains only contiguous allocations, for example the column $[1\ 0\ 1]^T$ is not contiguous allocation and cannot exist in $\mathbf{R}$.

Table 2.1: Summary of the most significant notation used in this chapter.

Symbol	Meaning
K, k	Number of UEs, UE Index
M	Number of RBs
J, j	Number of MCSs, MCS Index
$\mathbf{R}$	Contiguous allocation matrix
$\mathbf{r}_c$	The column number c in the matrix $\mathbf{R}$
c	RBs contiguous-chunk index
C	Total number of contiguous chunks possible
D_k^n	Maximum permitted-delay for bearer n of UE k
Q_k^n	Maximum queue length corresponding to D_k^n
$\bar{d}_k^n$	Average delay of bearer n of UE k (ms)
$\bar{q}_k^n$	Average queue length of bearer n of UE k (bits)
$\hat{q}_k^n$	Maximum allowed average queue length of bearer n of UE k
λ_k^n	Arrival rate of bearer n of UE k (ms)
$T_{k,j,c}$	TB size achieved by the choice k, j, c
β	Binary number indicator
P^{max}	Maximum transmission power threshold
E_k	TB size satisfies the rate and the delay constraints of UE k

The LTE frame duration is 10 ms and it is composed of 10 LTE subframes. Each subframe has a duration of 1 ms, and represents a TTI [9]. When the normal cyclic prefix is used [9], each subframe consists of 14 SC-FDMA symbols, each with a duration of $66.67\mu s$. Following the assumption used in [20], three symbols in each frame are assigned to uplink physical control signalling. The total number of data symbols in each RB per subframe is $(14-3)\times 12 = 132$. Fig. 2.1 shows the LTE subframe structure.

The LTE physical layer supports various modulation and coding schemes (MCSs). The MCS and the RBs that are assigned to a user determines the uplink transport block (TB) size. Suppose that the column vector $\mathbf{r}_c$ and the MCS number j are assigned to user k, the TB size that user k can transmit over a TTI is calculated as follows

$$T_k(j,c) = \left\lfloor 132 \times \zeta_j \times \|\mathbf{r}_c\| \right\rfloor \tag{2.4}$$

where ζ_j is the MCS efficiency for the MCS number j in terms of the number of useful bits per symbol, $\|\mathbf{r}_c\|$ is the the Hamming weight of $\mathbf{r}_c$, which represents the number of RBs that are allocated to user k, and $\lfloor x \rfloor$ denotes the largest integer number less than or equal to

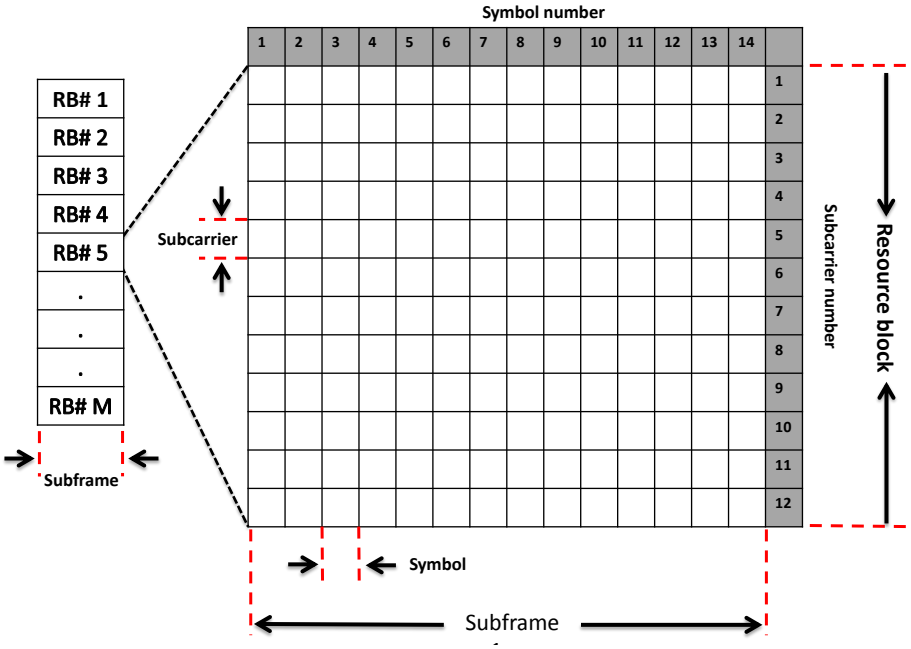

Figure 2.1: The structure of the LTE subframe.

x.

The selection of MCS is determined such that the BLER is lower than the target BLER, which is 10% for LTE systems [9]. The BLER for each MCS depends on the effective received signal to noise ratio (SNR). A simple criterion for choosing appropriate MCSs is based on a set of SNR thresholds [9]. Given the effective received SNR for a user, the eNB selects the most spectrally efficient MCS that satisfies BLER $< 10\%$. In practice, the BLER values are determined through a link-level simulator for all MCSs. Fig. 2.2 presents the simulation curves of the adopted MCSs in an LTE uplink over an additive white Gaussian noise (AWGN) channel [3]. Table 2.2 shows the modulation, code rate, spectral efficiency, and SNR thresholds for the MCSs [1]. The SNR thresholds can be obtained from Fig. 2.2.

The effective instantaneous received SNR for user k who is assigned column $\mathbf{r}_c$ at TTI t is given by [20, 23]

$$\gamma_k[t] = \frac{1}{\|\mathbf{r}_c\|} \sum_{m \in \mathbf{r}_c} \frac{\alpha_{k,m}[t] P_{k,m}[t]}{\sigma^2} \tag{2.5}$$

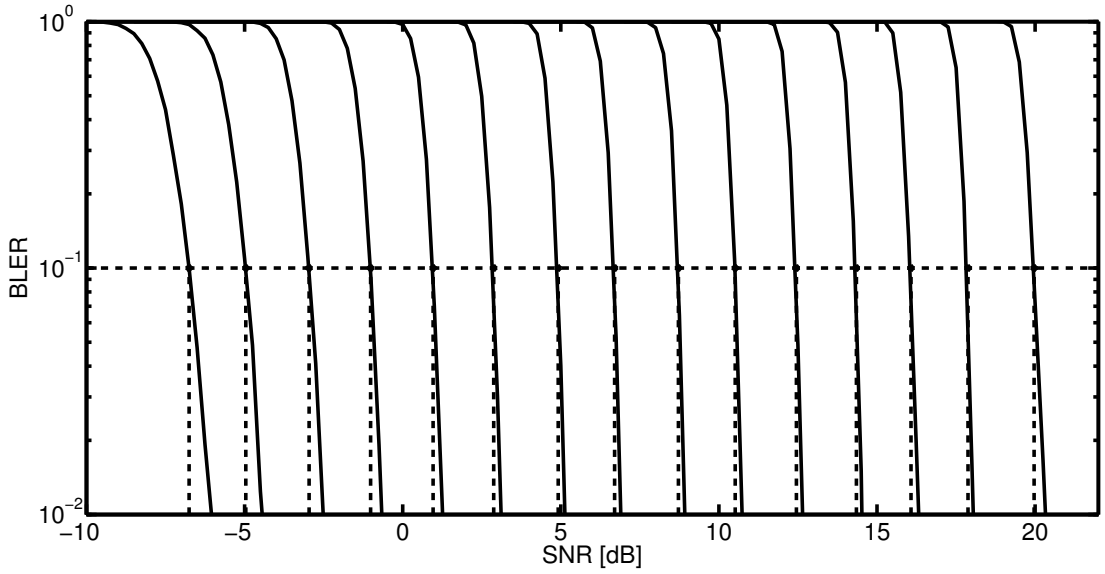

Figure 2.2: BLER-SNR curves for all Table 2.2 MSC, from Index 1 (leftmost) to index 15 (rightmost) [3].

where $P_{k,m}[t]$ is the transmission power of user k over the RB m during the TTI t, $\alpha_{k,m}[l] = \left|h_{k,m}[t]\right|^2$, $h_{k,m}[t]$ is the channel gain of RB m at TTI t seen by user k, and σ^2 is the AWGN variance.

To reduce the signalling overhead, LTE specifies that when a user is assigned more than one RB, one power level should be transmitted over the assigned RBs [15]. In other words $P_{k,1} = P_{k,2} = \cdots = P_{k,m}$. Therefore, the RB index m can be dropped, and (2.5) can be written as

$$\gamma_k[t] = \frac{1}{\|\mathbf{r}_c\|^2} \sum_{m \in \mathbf{r}_c} \frac{\alpha_{k,m}[t] P_k[t]}{\sigma^2} \tag{2.6}$$

where $P_k[t] = \|\mathbf{r}_c\| \, P_{k,m}[t]$ is the total transmission power of user k at TTI t. Using (2.6), the required power to achieve an effective SNR of $\gamma_k[t]$ is

$$P_k[t] = \gamma_k[t] \left(\sum_{m \in \mathbf{r}_c} \alpha_{k,m}[t] \right)^{-1} \|\mathbf{r}_c\|^2 \sigma^2 \tag{2.7}$$

To recap, when user k transmits P_k over the RB chunk that is represented by $\mathbf{r}_c$, the effective received SNR of the user is γ_k, which can be mapped to an appropriate MCS

Table 2.2: List of MCS indices [1]

Index j	Modulation	Coding Rate	Spectral Efficiency ζ	Effective SNR (dB) γ_k
0	—	—	0 bits	> -6.7536
1	QPSK	78/1024	0.15237	$-6.7536 : -4.9620$
2	QPSK	120/1024	0.2344	$-4.9620 : -2.9601$
3	QPSK	193/1024	0.3770	$-2.9601 : -1.0135$
4	QPSK	308/1024	0.6016	$-1.0135 : +0.9638$
5	QPSK	449/1024	0.8770	$+0.9638 : +2.8801$
6	QPSK	602/1024	1.1758	$+2.8801 : +4.9185$
7	16QAM	378/1024	1.4766	$+4.9185 : +6.7005$
8	16QAM	490/1024	1.9141	$+6.7005 : +8.7198$
9	16QAM	616/1024	2.4063	$+8.7198 : +10.515$
10	64QAM	466/1024	2.7305	$+10.515 : +12.450$
11	64QAM	567/1024	3.3223	$+12.450 : +14.348$
12	64QAM	666/1024	3.9023	$+14.348 : +16.074$
13	64QAM	772/1024	4.5234	$+16.074 : +17.877$
14	64QAM	873/1024	5.1152	$+17.877 : +19.968$
15	64QAM	948/1024	5.5547	$> +19.968$

through Table 2.2. Consequently, the TB size of user k can be calculated from (2.4).

2.3.1 LTE QoS and Buffer Status Reports (BSRs)

LTE systems are designed to support a wide range of applications and services. In general, the user might run multiple applications simultaneously, each application requires different QoS. For example, a user can play real-time game while downloading a file using a file-transfer protocol. The eNB establishes multiple radio bearers per user to support multiple QoS requirements as shown in Fig. 2.3. The LTE defines two main radio bearer categories [9]:

1. Bearers with guaranteed bit rate (GBR) are established for real-time applications such as voice and video, which require certain GBR to satisfy their QoS requirements.
2. Non-GBR (NGBR) bearers do not guarantee any particular bit rate and are used for non-real-time applications such as buffered video streaming.

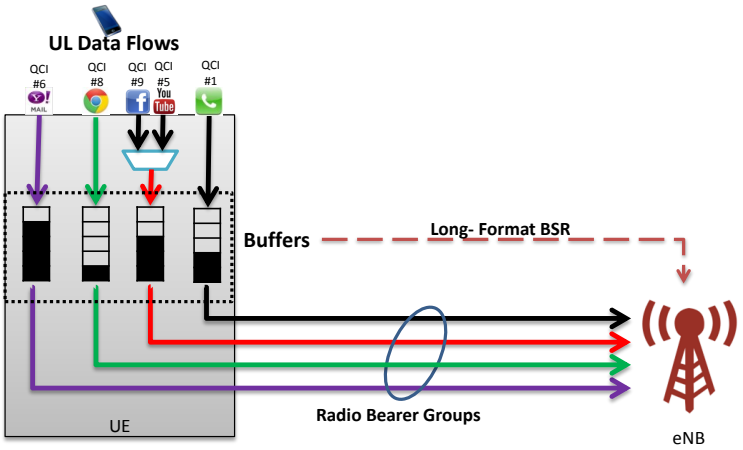

Figure 2.3: An example of four bearers established for a user.

The QoS class identifier (QCI) and the allocation and retention priority (ARP) are the bearer QoS parameters. The QCI is a scalar that specifies Internet Protocol (IP) level packet characteristics of the bearers. The ARP of a bearer is used for admission and handover control. When the bearer is established, the ARP has no effect on packet transmissions, which are managed only by their QCI specifications [9]. LTE uplink scheduling takes place in the eNB, where information about buffered data sizes is reported for all UEs who have data to be transmitted using the buffer status reporting (BSR) mechanism. A BSR has two possible formats [9]: short BSR format and long BSR format. The short format reports one bearer, while the long format reports up to four bearers. The choice of short or long format is determined by the number of non-empty buffers. If a user has only one bearer, the short format is used to conserve channel resources because the short format report requires fewer bits. Although UEs may have more than four non-empty buffers, the maximum number of reporting bearers is four. In the long format scheme, bearers are grouped into four groups before they are reported, and therefore considering users with four bearers is practically acceptable. In this chapter, it is assumed that each user has a maximum of four bearers, where each bearer is modeled as an infinite first-in first-out buffer in the radio link control (RLC) sub-layer.

2.3.2 Uplink Data Transmission Procedure

The sequence of uplink data transmission is shown in Fig. 2.4. The UE receives uplink traffic from upper layers. Data for multiple logical channels are queued in the RLC sub-layer buffers. Information about buffered data sizes is reported to the eNB over the physical uplink control channel using the BSR procedure. The scheduler in the eNB makes allocation decisions according to a specific scheduling policy. Based on the allocation decisions, the eNB sends allocation maps to the users over the physical downlink control channel. The user's allocation map specifies the assigned RBs, power control entity and MCS [24]. The power control entity specifies the uplink transmission power for each user. The RBs chunk and MCS that is assigned to a user determine the uplink transport block TB size. However, how the TB is shared between users' buffers is left to the user policy, which is discussed in Section 2.8.

In the UE media access control (MAC) sub-layer, a MAC protocol data units (PDU) is formed according to the received map allocation. The MAC PDU contains data from one or more RLC PDUs in addition to the MAC header. The MAC passes the MAC PDU to the physical layer (PHY), which adds the cyclic redundancy check (CRC) bits to the MAC PDU and then transmits the entire packet as a TB over the physical uplink shared channel to the eNB. The PHY responsibility is to deliver the TB with an error probability less than a targeted BLER.

2.3.3 Delay Analysis

As mentioned in Section 2.3.1, the BSR does not report explicit information about PDU delays, but reports the sizes of the queued data in the UE buffers. In this section, the PDU delay is mapped to the size of the queued data. The queue evolution during TTI $t+1$ can be described as

$$q_k^n[t+1] = q_k^n[t] + a_k^n[t] - T_k^n[t] \tag{2.8}$$

where $q_k^n[t]$ is the number of pending bits at the beginning of TTI t, $a_k^n[t]$ is the number of bits arriving at TTI t with an average arrival rate of λ_k^n, and $T_k^n[t]$ is the number of transmitted bits from bearer n of user k at TTI t.

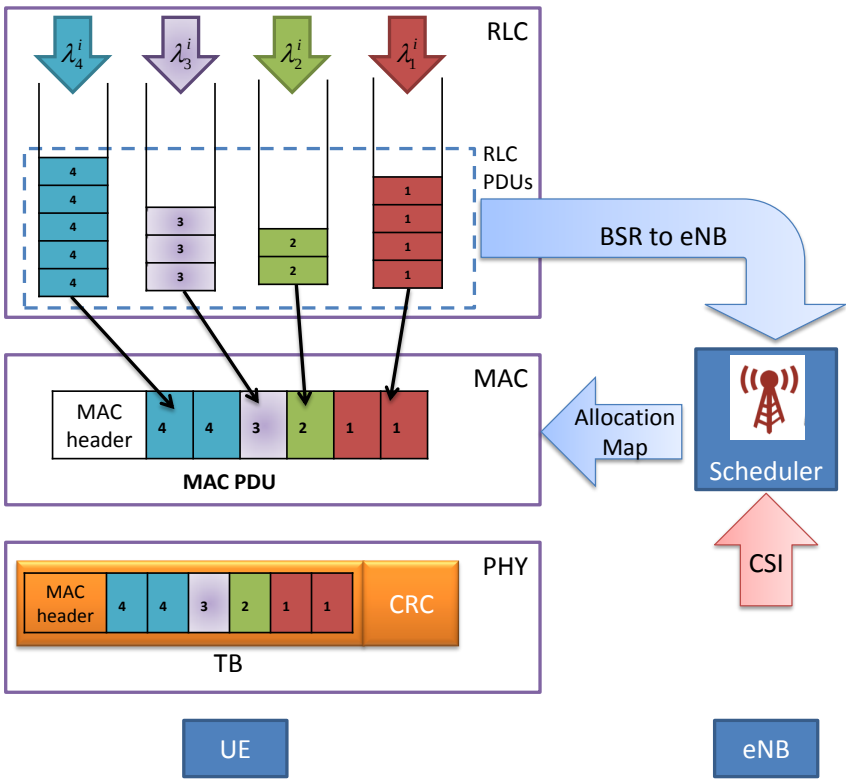

Figure 2.4: Uplink data transmission sequence.

LTE defines a packet delay budget (PDB) for each bearer, which defines an upper bound for the time that a packet may be delayed between the UE and the packet data network gateway. In LTE, the outage delay outage level should be less than 2% [9]

$$P(d_k^n > \check{D}_k^n) < 0.02 \tag{2.9}$$

where d_k^n is the head of the line packet delay, and $\check{D}_k^n$ is the PDB. For high data traffic load, the following approximation is valid [25]

$$P(d_k^n > \check{D}_k^n) \approx e^{\left(-\check{D}_k^n/\bar{d}_k^n\right)} \tag{2.10}$$

where $\bar{d}_k^n$ is the average value of d_k^n. Therefore, in the case of high data traffic load, controlling $\bar{d}_k^n$ is approximately equivalent to controlling the delay violation probability $P(d_k^n > \check{D}_k^n)$. The justification for the assumption of heavy data traffic is that most delay violations take place when the traffic load is heavy. By substituting (2.9) into (2.10), the required average delay for bearer n of user k is given by

$$\bar{d}_k^n < -\frac{\check{D}_k^n}{\ln(0.02)}. \tag{2.11}$$

This work focuses on the the air-interface delay between UEs and eNB. It is assumed that the air-interface delay D_k^n contributes to δ of the PDB as follows

$$D_k^n = \delta\ \check{D}_k^n \tag{2.12}$$

where $0 \leq \delta \leq 1$. Using Little's theorem [22], the average delay can be computed as

$$\bar{d}_k^n = \frac{\bar{q}_k^n}{\lambda_k^n} \tag{2.13}$$

where $\bar{q}_k^n$ is the average queue size. Substituting (2.13) and (2.12) into (2.11), the bearer air-interface delay can be controlled by controlling the average number of bits in the users' queues

$$\bar{q}_k^n < \hat{q}_k^n \tag{2.14a}$$

$$\hat{q}_k^n = -\frac{Q_k^n}{\ln(0.02)} \tag{2.14b}$$

where $\hat{q}_k^n$ is the maximum allowable average queue length of bearer n of user k that satisfies the delay requirement, and $Q_k^n = D_k^n \times \lambda_k^n$ is the average queue length that corresponds to the air-interface delay D_k^n.

It is worth noting that the size of the data buffers is assumed to be greater than the PDBs for all bearers. If the size of a buffer is less than the PDB, the PDB is assumed to be the maximum buffer size for that bearer.

It is assumed that the eNB knows, or at least can estimate, the average arrival rates for all bearers and users. One way to estimate the average arrival rates is reported in [26]. Suppose that the scheduler successfully maintains the delay requirements of buffer number n of user k such that $q_k^n \leq \hat{q}_k^n$, the average arrival rate λ_k^n is equal to the long-term average of the service rate T_k^n.

2.4 System Constraints and Objective

The resource allocation in LTE uplink requires the following constraints to maintain the physical layer restrictions and the QoS requirements:

1. Exclusivity constraint: for every TTI, a single RB is allocated to no more than one user.
2. Contiguity constraint: SC-FDMA restricts the RB allocations to be contiguous. Each column in the matrix $\mathbf{R}$ represents a contiguous RB allocation. The contiguity constraint can be maintained by assigning one column from the matrix $\mathbf{R}$ to each user.
3. Power constraint: the LTE standard specifies P^{max} as the maximum transmission power threshold that the user cannot exceed.
4. Rate constraint: minimum bit rate for the GBR bearers must be maintained.
5. Delay constraint: NGBR bearers subject to PDB. As discussed in Section 2.3.3, the delay constraint can be maintained by controlling the average queue length.

The objective of this work is to minimize the sum of transmission power consumed by the users while maintaining their QoS. This chapter is seeking to answer the following

question: how should the available resources, in terms of RBs and transmission power, be assigned to the users so that the total transmission power is minimized without violating the users' QoS requirements? Without loss of generality, a finite time horizon of length F TTIs is chosen. It is assumed that the current time is t and the observation interval is $[t, t+F]$. Denote $W(x[t], L)$ as a sliding-average window of length L for variable x

$$W(x[t], L) = \frac{1}{L} \sum_{l=t}^{t+L} x[l], \tag{2.15}$$

and denote $P_{k,j,c}[l]$ as the transmission power required to achieved the target BLER when user k transmits using the MCS number j over the RB allocation number c at TTI l. The resource allocation problem can be written as

$$\min \sum_{l=t}^{t+F} \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{c=1}^{C} P_{k,j,c}[l] \beta_{k,j,c}[l] \tag{2.16a}$$

subject to

$$\bigcap_{k=1}^{K} \mathbf{r}_c \, \beta_{k,j,c}[l] = \emptyset, \ \forall l, j, c \tag{2.16b}$$

$$P_{k,j,c}[l] \leq P^{max}, \forall l, k, j, c \tag{2.16c}$$

$$W(T_k^n[t], F) \geq r_k^n \tag{2.16d}$$

$$W(q_k^n[t], F) \leq D_{marg} \times \hat{q}_k^n \tag{2.16e}$$

where $\beta_{k,j,c}[l]$ is a binary number indicator that is equal to 1 if and only if the MCS number j and the column $\mathbf{r}_c$ are selected for user k at TTI l, J is the total number of MCSs, $D_{marg} \in (0, 1)$ is a margin used to maintain the delay less than the maximum in (2.14), and r_k^n is GBR of the bearer number n of user k. Note that the exclusivity constraint is maintained by (2.16b), where the summation over c restricts the users to only contiguous RB allocations and maintains the contiguity constraint. The power, rate, and delay constraints are maintained by (2.16c), (2.16d), (2.16e), respectively.

The problem shown in (2.16) is a discrete time stochastic control process. One way to solve this problem is to formulate it as a constrained MDP. Although general techniques

exist to solve MDPs, they suffer from the curse-of-dimensionality problem [27], where the number of states grows vastly with the numbers of both users and bearers. Consequently, formulating and solving constrained MDPs is non-trivial as it deals with an extremely large number of transition probabilities. Therefore, approximated solutions are often provided [27]. Moreover, a comprehensive knowledge of the users' channel gains and arrival processes may be required to solve the MDP. A similar problem in OFDM systems is investigated in [12] by relying on stochastic convex optimization. For each TTI, the allocation depends on the instantaneous channel gains and Lagrange multipliers. The multipliers are associated with the QoS requirements and are estimated online using stochastic approximation tools. However, many approximations are used while turning this problem into a stochastic convex optimization. First, Shannon's capacity formula is used (continuous rate adaptation) instead of a practical discrete set of MCSs. Second, the exclusivity constraint is relaxed. The relaxation replaces the binary indicators by contentious variables belong to the interval $[0, 1]$. In OFDMA each binary number refers to a single RB or subcarrier. Having a fraction of a subcarrier translates into time sharing between users for the subcarrier which creates a form of time-division multiple access. However, in SC-FDMA, the allocated RBs for a particular user should be adjacent, and sharing chunks of RBs over time is not applicable because the contiguous allocation is not guaranteed [13]. Third, it is observed that, optimum values of the multipliers can never be found. Therefore stochastic approximation is used to estimate the multiplier values.

An alternative approach to solve (2.16) is to design a non-causal optimal offline scheduler, which gives a guideline to design and evaluate online suboptimal schedulers. The offline optimal scheduler requires a prior knowledge of both arrival data units and channel state information for $\{t, t+1, ..., t+F\}$ at TTI t. Therefore, the problem turns into a discrete time deterministic control process, which can be solved, for example, by binary integer programming. Nevertheless, the search space of this formulation is extremely large and can be calculated as follows. For one TTI, for user k, there are $\frac{1}{2}M(M+1)$ possible allocations in the frequency domain as shown in (2.2). For each possible allocation, J MCSs exist. Therefore, the total number of choices possible for user k is $\frac{J}{2}M(M+1)$. For the K users, the search space is $\left(\frac{J}{2}M(M+1)\right)^K$. Consequently, the search space for F

TTIs is

$$\left(\frac{J}{2}\,M(M+1)\right)^{KF}. \tag{2.17}$$

For example, assume a system with the following parameters: $M = 6$ RBs, $J = 15$ MCSs, $K = 4$ and $F = 10$ TTIs. The search space size is 8.5590×10^{99}.

In the following sections, a computationally feasible version of the scheduling problem is presented.

2.5 BIP Formulation

To simplify the scheduler and to avoid the computationally-excessive optimization in (2.16), a formulation of the optimization problem for a single time slot ($F = 1$) is presented here. The exclusivity, contiguity, and power constraints must be satisfied for each TTI because they are related to the physical system. However, the rate and delay constraints are based on averages, which means they are soft and it is not necessary to satisfy them every TTI. In other words, the rate and/or delay might not meet their QoS requirements at certain times, but on average over a long time interval, the QoS requirements are met. Allowing the rate and delay constraints to be soft avoids infeasible instantaneous solutions. Such solutions appear when the instantaneous required data transmission rate is greater than the instant channel capacity. However, it is assumed that, in the long term, the channel capacity can provide the required QoS. In cases where the QoS requirements are greater than the available channel capacity, scheduling becomes infeasible, and dropping users may be the only feasible solution [12]. An admission control procedure is responsible for deciding which user should be dropped or admitted in such cases. This study does not consider admission control procedures, and it is assumed that the average channel capacity can manage the required QoS.

The following key observation can be extracted from (2.7). Given that user k has to transmit T_k bits, the transmission power of user k can be decreased by a) increasing the effective SNR by assigning RBs that have less fading; and/or b) transmitting the T_k bits over a longer period of time (more subframes). The second point needs more elaboration. Fig. 2.5 shows a logarithmic relationship between the spectral efficiency of the MCSs in

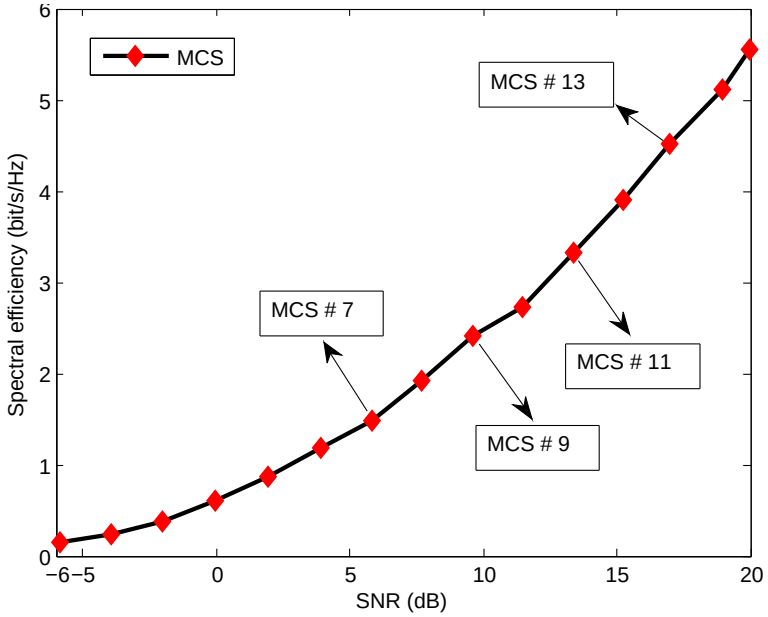

Figure 2.5: Spectral efficiency versus SNR for the MCS that are shown in Table 2.2.

Table 2.2 and their corresponding effective SNRs. From (2.7), the transmission power decreases as a function of the effective SNR for a specific RB chunk allocation, which implies that the transmission power is logarithmically related to the spectral efficiency of the transmission. For example, transmitting four symbols using MCS number four is equivalent to transmitting one symbol using MCS number nine. Although, the former transmission requires additional four time slots, it consumes only 43% of the later transmission power consumption. Therefore, splitting the data and transmitting lower data rates over more subframes eventually lowers the total transmit power. However, the data rates should maintain the users' QoS requirements.

Modulation and coding schemes are less power-efficient at higher transmission rates [28]. Therefore, the scheduler is designed to judiciously transmit the minimum number of bits that softly satisfies the rate and the delay constraints at each TTI. Then by optimal power allocation, optimal chunk of RBs is assigned to each user. Define B_k^n as the minimum number of transmit bits that satisfy both the rate and the delay constraints to bearer n of user k at TTI t. Therefore, the total TB for user k that satisfies the rate and the delay

constraints is

$$E_k = \sum_{n=1}^{4} B_k^n.$$

As the rate and the delay constraints are converted to soft constraints, the minimum-cost solution is a trivial one, i.e., no power is consumed if no data is transmitted. To address this issue, an extra weight $\rho_{k,j,c}$ is defined and added to the applied power cost as follows

$$\rho_{k,j,c} = \log(E_k) - \log(T_{k,j,c}/E_k^{GBR}), \;\; T_{k,j,c} > 0, \tag{2.18}$$

where the $T_{k,j,c}$ are the TB achieved using MCS j over the RB allocation number c, and E_k^{GBR} is the number of bits that satisfies all the GBR bearers of user k

$$E_k^{GBR} = \sum_{n \in \mathcal{N}_k^{GBR}} B_k^n, \tag{2.19}$$

where $\mathcal{N}_k^{GBR}$ is a set contains the index of the GBR bearers that belongs to user k.

The weights $\rho_{k,j,c}$ measure how close the E_k are to the actual transmitted TB $T_{k,j,c}$. To better illustrate extra weights, consider the demonstration shown in Fig. 2.6. Two main interesting characteristics can be observed. First, users with higher E_k^{GBR} values have higher weights, which gives them a higher priority through the scheduling. Second, as the number of bits transmitted increases, weight values drop rapidly when $T_{k,j,c} < E_k^{GBR}$, but slightly when $T_{k,j,c} > E_k^{GBR}$. The second characteristic implies that satisfying GBR requirements for all users are more important than that for NGBR requirements.

In this context, for single TTI, (2.16) is expressed as (time index is omitted for simplicity)

$$\min \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{c=1}^{C} \left(\rho_{k,j,c} + P_{k,j,c}\right) \beta_{k,j,c} \tag{2.20a}$$

subject to

$$T_{k,j,c} \leq E_k \tag{2.20b}$$

$$\text{(2.16b), (2.16c)} \tag{2.20c}$$

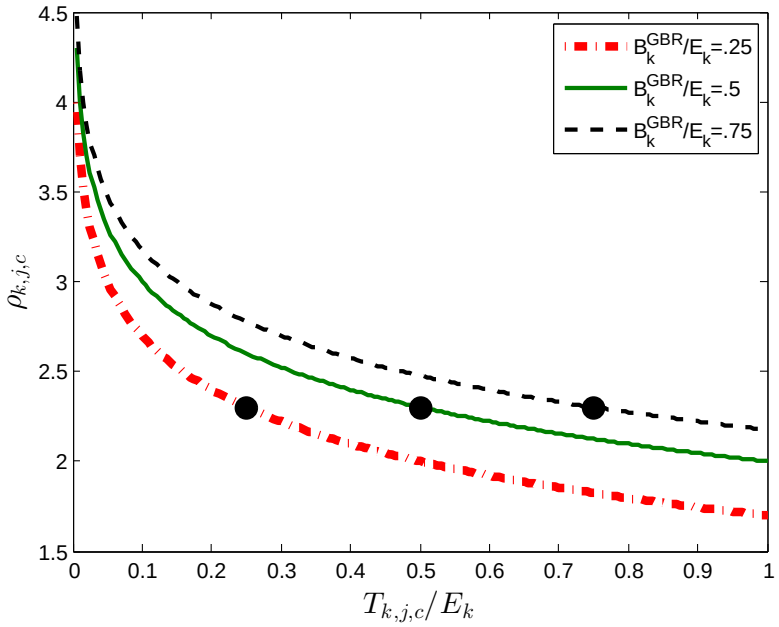

Figure 2.6: Extra weight demonstration for $E_k = 200$ bits.

2.5.1 Binary Integer Programming

The *bintprog* optimization package from MATLAB is used to solve the non-convex problem. The general form of a BIP optimization can be presented as

$$\min \mathbf{c}^T\mathbf{x} \tag{2.21a}$$

subject to

$$\mathbf{A}\mathbf{x} \leq \mathbf{b}, \quad \mathbf{A}_{eq}\mathbf{x} = \mathbf{b}_{eq}, \tag{2.21b}$$

where the vector $\mathbf{c}$ represents the weights $\rho_{k,j,c}$ plus the transmission power costs $P_{k,j,c}$, and the binary decision vector $\mathbf{x}$ minimizes the objective function and represents the term $\beta_{k,j,c}$ in (2.20a). Constraints (2.20b), (2.20c) are maintained by linear inequality and equality constraints, $\mathbf{A}\mathbf{x} \leq \mathbf{b}$, $\mathbf{A}_{eq}\mathbf{x} = \mathbf{b}_{eq}$, respectively, where $\mathbf{A}$, $\mathbf{A}_{eq}$ are matrices containing the coefficients of the inequality and equality constraints, and $\mathbf{b}$, $\mathbf{b}_{eq}$ are vectors that fulfill the inequality and equality constraints, respectively. The exclusivity and contiguity constraints are defined as follows. Consider user k with contiguous allocation matrix $\mathbf{R}$ as in (2.3) with a dimension of $M \times C$. Each column in $\mathbf{R}$ presents a feasible contiguous allo-

cation. For each feasible contiguous allocation, J different MCSs are possible, the matrix which contains all feasible contiguous allocations for all MCSs is defined as

$$\mathbf{A}_k = [\underbrace{\mathbf{R}, \mathbf{R}, \ldots, \mathbf{R}}_{J}]. \tag{2.22}$$

Each column in (2.22) is associated with a potential transmit power cost calculated from $P_{k,j,c}$. The power constraint is maintained by deleting columns with potential transmission power greater than P^{max}. Define the matrix $\mathbf{A}_k^{th}$ which has all columns in $\mathbf{A}_k$ less than or equal to P^{max}. Define matrix $\mathbf{A}$ which concatenates all matrices $\mathbf{A}_k^{th}$ for all users as follows:

$$\mathbf{A} = \left[\mathbf{A}_1^{th}, \mathbf{A}_2^{th}, \ldots \mathbf{A}_K^{th}\right], \tag{2.23}$$

where each row in $\mathbf{A}$ represents a single RB, to ensure the exclusivity constraint, and the vector $\mathbf{b} = \mathbf{1}_M$, where $\mathbf{1}_M$ is a vector of M ones. The equality constraints maintain a unique selection from all feasible allocations for all data sizes per user, and therefore

$$\sum_{j=1}^{J} \sum_{c=1}^{C_k} \beta_{k,j,c} = 1, \forall k.$$

The equality constraints are expressed as follows

$$\mathbf{A}_{eq} = \begin{bmatrix} 1^T_{\left|\mathbf{A}_1^{th}\right|} & \cdots & 0^T_{\left|\mathbf{A}_K^{th}\right|} \\ \vdots & \ddots & \vdots \\ 0^T_{\left|\mathbf{A}_1^{th}\right|} & \cdots & 1^T_{\left|\mathbf{A}_K^{th}\right|} \end{bmatrix} \tag{2.24}$$

where $\left|\mathbf{A}_k^{th}\right|$ is the number of columns in $\mathbf{A}_k^{th}$ which denotes the number of potential allocation choices for user k, and $\mathbf{1}_x^T, \mathbf{0}_x^T$ are row vectors of length x of ones and zeros, respectively. The vector $\mathbf{b}_{eq}$ is defined as $\mathbf{b}_{eq} = \mathbf{1}_K$ to guarantee that only one of the possible allocations is assigned for each user.

2.5.2 Complexity of BIP

Consider the worst case scenario, where all columns in $\mathbf{A'}_k, \forall k$ are less or equal to P^{max}. The search space of BIP formulation is similar to the search space of (2.17) for a single TTI ($F = 1$)

$$\left(\frac{1}{2} J\, M(M+1)\right)^K. \tag{2.25}$$

For example, assume a system with the following parameters: $M = 6$ RBs, $J = 15$ MCSs, and $K = 4$. The BIP worst case search space size is 9.8456×10^9. Thus, the approach as formulated is still computationally expensive. Therefore, low-complexity algorithms are needed to solved the resource allocation problem.

2.6 Iterative Algorithm

In this section, an iterative algorithm is proposed to solve the BIP with much less computational complexity. The algorithm belongs to the greedy algorithm family. The objective is to minimize the summation of the total users' costs by assigning RBs to the users iteratively. In each iteration, a single RB is assigned to a user who can achieve maximum reduction in the cost function. Therefore, the algorithm needs M iterations to assign all the RBs to users. For each iteration, the algorithm finds the best RB for each user. The best RB of user k is defined as the RB that has the highest instantaneous channel gain ($\alpha_{k,m}$) and satisfies the contiguous allocation of user k. Then, the change in the user's cost value before and after assigning the best RB is calculated for each user. As the algorithm is greedy, the user who has the maximum positive change in the user's cost function is granted the allocation. The pseudo-code in Table 2.3 describes the algorithm. The proposed algorithm consists of four main steps as follows:

Lines 1-7: find the minimum rate (E_k) for each user k that satisfies the rate and the delay constraints. In addition, in this step the following parameters are initialized: the set of users index $\mathcal{K}$, the set of non-assigned RBs $\mathcal{M}$, the set of RBs that assigned to users $\mathcal{M}_k, \forall k \in \mathcal{K}$, and the initial cost for each user $\Psi_k = \log(E_k), \forall k \in \mathcal{K}$. It is worth noting that the initial cost of a user is equal to maximum extra weight of the user.

Lines 8-14: find the best feasible RB $\mathcal{F}_k$ for each user k. Users, who have been assigned one or more RB, are limited to their neighboring RBs (line: 11). Therefore, two RBs are feasible at most to any user who has been assigned one or more RBs. For example, suppose that user k has been assigned the RB numbers $\{4, 5, 6\}$. Therefore, the feasible RBs to user k are 3 and 7 if they are not assigned to any user. However, users who have not yet been assigned RBs can select any RB of the non-assigned RBs set $\mathcal{M}$ (line: 13). In the case that a user has more than one feasible RB, the algorithm selects the RB that has the maximum instantaneous channel gain.

Lines 15-21: after finding the best feasible RB to each user, the users' cost values of the new potential allocation sets are computed $\Psi_k^*, \forall k \in \mathcal{K}$. Then, the potential cost reduction $\Delta\Psi_k$ for each user is computed, where $\Delta\Psi_k$ is the difference between the actual cost value Ψ_k and the potential cost value Ψ_k^*. If all users do not benefit (reduce their cost values) from the new potential allocation sets, the algorithm stops (lines 19 and 20). The $\mathcal{COST}(\mathcal{M}_k, E_k)$ function finds the total costs, i.e. $\rho_{k,j,c}+P_{k,j,c}$, associated with assigning the set of RBs $\mathcal{M}_k$ to user k. In case of $\mathcal{M}_k = \emptyset$, the total cost is $\mathcal{COST}(\emptyset, E_k) = \log(E_k)$, as illustrated in line 5.

Lines 22-26: determine the winning user who achieved the maximum cost-reduction $\Delta\Psi_k^*$ (lines 22). Then the algorithm assigns RB $\mathcal{F}_k^*$ to the winning user (lines 24), and updates the set of non-assigned RBs (lines 25).

2.6.1 Complexity of the Iterative Algorithm

For each major iteration (lines 8-27), an RB is allocated, and therefore, the maximum number of major iterations is M. The first major iteration for each user requires at most M operations to compute costs. When a user is assigned one or more RB, this number is reduced to two at most to the lower and upper RB. Assume that each user performs two operations. For K users there are $2 \times K$ operations in each major iteration. Therefore, the complexity of the proposed iterative algorithm is $O(MK)$. In a similar example to that in Section 2.5.2, the complexity of the worst case is in the order of 24 iterations, which is significantly less than the optimal algorithm complexity. For a realistic scenario,

Table 2.3: Iterative allocation

$\mathcal{M} = \{1, 2, \ldots, M\}$
$\mathcal{K} = \{1, 2, \ldots, K\}$
for $k \in \mathcal{K}$ **do**
  find E_k
  $\Psi_k = \log(E_k)$
  $\mathcal{M}_k = \emptyset$
end for
while $|\mathcal{M} \neq \emptyset|$ **do**
  for $k \in \mathcal{K}$ **do**
    if $\mathcal{M}_k \neq \emptyset$ **then**
      $\mathcal{F}_k = \underset{m \in \{\min(\mathcal{M}_k)-1, \max(\mathcal{M}_k)+1\} \cap \mathcal{M}}{\arg\max} \{\alpha_{k,m}\}$
    else
      $\mathcal{F}_k = \underset{m \in \mathcal{M}}{\arg\max} \{\alpha_{k,m}\}$
    end if
    $\mathcal{M}_k^* = \mathcal{M}_k \cup \mathcal{F}_k$
    $\Psi_k^* = \mathcal{COST}(\mathcal{M}_k^*, E_k)$
    $\Delta\Psi_k = \Psi_k - \Psi_k^*$
  end for
  if $\Psi_k^* < 0, \ \forall k \in \mathcal{K}$ **then**
    exit
  else
    $k^* = \underset{k}{\arg\max} \{\Delta\Psi_k\}$
    $\Psi_k = \Psi_k^*$
    $\mathcal{M}_k^* = \mathcal{M}_k \cup \mathcal{F}_k$
    $\mathcal{M}^* = \mathcal{M} \setminus \mathcal{F}_k$
  end if
end while

the maximum number of RBs in LTE system is 100 RBs. Thus, the complexity of such scenario is in the order of $10,000$ iterations.

2.7 Proportional Fair Scheduler

The PF scheduler has been widely investigated in the literature. The objective of the PF scheduler is maximizing the total throughput of the system while maintaining level of fairness between users. Lee *et al.* [19] investigated the PF scheduler for SC-FDMA systems. The scheduling problem is known to be an NP-hard.

To determine the power efficiency of the proposed algorithms, the PF scheduler is used as the baseline scheme for comparison. The PF scheduler of [19] have been modified to cope with our system model. The objective function at TTI t can be expressed as

$$\max \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{c=1}^{C_k} \omega_k[t] \times T_{k,j,c}[t], \tag{2.26a}$$

subject to

$$\text{(2.16b), (2.16c)} \tag{2.26b}$$

where $\omega_k[t]$ is a scheduling weight assigned to user k at TTI t. The scheduling weights depend on transmission history for users as follows

$$\omega_k[t] = \frac{1}{W(T_k[t-1], L_{PF})} \tag{2.27}$$

where $W(T_k[t-1], L_{PF})$ is a sliding-average window defined in (2.15) of length L_{PF}. Users who have low historical average data rates are assigned higher weights than those who have high historical average data rates, which increases their chances of obtaining more RBs during the scheduling.

The PF scheduling worst-case search space is the same as BIP. However, for the BIP scheduler, the number of available MCSs is often less than that for PF scheduler because the BIP scheduler avoids rate transmission higher than E_k. Therefore, the complexity of the BIP is expected to be lower than that for the PF.

Table 2.4: Intra-user scheduling for user k

$E_k^{GBR} = \sum_{n \in \mathcal{N}_k^{GBR}} B_k^n$
$E_k^{NGBR} = \sum_{n \notin \mathcal{N}_k^{GBR}} B_k^n$
if $E_k^{GBR} < T_k$ **then**
  $T_k^n = B_k^n,\ \forall n \in \mathcal{N}_k^{GBR}$
  $T_k^n = (T_k - E_k^{GBR}) \times (B_k^n / E_k^{NGBR}),\ \forall n \notin \mathcal{N}_k^{GBR}$
else
  $T_k^n = T_k \times (B_k^n / E_k^{GBR}),\ \forall n \in \mathcal{N}_k^{GBR}$
  $T_k^n = 0,\ \forall n \notin \mathcal{N}_k^{GBR}$
end if

2.8 Intra-User Scheduling

In all the scheduling schemes discussed above, the scheduler output is the set $T_k,\ \forall k = 1, 2, ..K$, which indicates the TB size for each user. It is assumed that UEs share their TBs between their bearers as follows:

1. GBR bearers should be satisfied before NGBR bearers.
2. Within the same radio bearer category (GBR or NGBR), the allocated resources are distributed proportionally to B_k^n. In case of PF scheduler, B_k^n are replaced by the queues length of bearer n.

The pseudo-code in Table 2.4 describes the intra-user scheduling, where T_k^n denotes the portion of T_k allocated to bearer n of user k.

2.9 Numerical Results

The simulation default parameters are shown in Table 2.5. The channel is modeled as a quasi-static frequency-flat Rayleigh fading channel. The channel gain is assumed to be constant over each RB bandwidth, but change independently over consecutive RBs. Moreover, it is assumed that users experience independent fading. For a Rayleigh fading channel, the distribution of the instantaneous received channel gain α follows the exponential distribution [29]

$$p\left(\alpha_{k,m}\right) = \frac{1}{\alpha_k} e^{-\frac{\alpha_{k,m}}{\alpha_k}} \tag{2.28}$$

Table 2.5: Simulation default parameters

Parameter	Value	Parameter	Value
Coherence time	1 ms	M	10
UE #	2	Iteration #	1e4
BW	15 kHz	Cells interference	Avoidance
δ	1	D_{marg}	0.9
P^{max}	23 dBm	L_{PF}	50
Channel fading	Rayleigh	L	8
Target BLER	10 %	σ^2	1

Table 2.6: Users' data profile

Index	Type	λ_k^n(Kbps)	D_k^n(ms)	Q_k^n(bits)	$\hat{q_k^n}$	r_k^n(Kbps)
$B1$	NGBR	100	20	2000	511	0
$B2$	GBR		-	-	-	50
$B3$	NGBR	75	20	1500	383	0
$B4$	GBR	-	-	-	-	30

where α_k is the expected value of $\alpha_{k,m}$ and denotes the average channel gain (ACG).

Table 2.6 presents bearers profiles. NGBR bearers serve non real-time applications. It is assumed that the data arrivals for non real-time traffic follow a Poisson distribution as it has been used in related works [22, 26, 30].

GBR bearers serve real time applications which require minimum guaranteed bit rates. Transmission resources are reserved for GBR bearers in the admission control function. Therefore, an eNB establishes GBR bearers "on demand" for users [31]. In this chapter, the virtual token queue (VTQ) is adopted to model the GBR buffers as used in [32, 33]. Each bearer is modelled as a VTQ, where tokens arrive at a constant rate equal to the guaranteed bit rate of the bearer. In other words, for each TTI a token of size r_k^n is added to the VTQ corresponding to queue n of user k. The number of tokens is reduced according to the actual amount of data transmitted for the bearer. The minimum number of bits that satisfies a GBR bearer is determined by the number of tokens in the virtual queue. For example, assume the virtual queue has $q_k^n[t]$ tokens, the minimum bit rate that satisfies the bit rate requirements of the bearer n of user k is $r_k^n\ q_k^n[t]$ bits.

2.9.1 Experiment 1: Two Users with Identical Conditions

In this experiment, a two-user scenario with an identical traffic load and channel profile is considered. Because the users experience identical conditions in terms of channel and traffic load, it is sufficient to show results for only one user. Each user is assumed to have two bearers, namely $B1$ and $B2$.

Fig. 2.7 compares the transmit power for the three schedulers. The BIP consumes slightly less power than the iterative scheduler, and both of them consume approximately 43% less power than the PF scheduler. To better visualize the efficiency of the schedulers, consider the following scenario: users transmit data until the total power consumption reaches 80 Watt. Note that 80 Watt is equivalent to transmitting at the maximum power threshold (200 mW) for 400 TTIs. Fig. 2.8 shows that the BIP and iterative schedulers prolong the battery life considerably compared with the PF scheduler for the same QoS requirements.

Fig. 2.9 compares the average queue length of the two bearers for the three schedulers. All the schedulers succeed in maintaining average queue lengths of the NGBR data less than the threshold. As the intra-scheduling serves the GBR bearers before the NGBR bearers, the queue lengths of the GBR bearers are shorter than the queue lengths of the NGBR for all schedulers. The PF scheduler tends to transmit aggressively, i.e., the maximum achievable data rate. Therefore, the average queue lengths of the PF scheduler are lower than those for the other two schedulers. The queue lengths of the BIP scheduler are slightly shorter than the queue lengths of the iterative scheduler.

Fig. 2.10 shows the probability density function (PDF) of the delay for the NGBR bearer at average channel gain of 10. Note that delay violations occur when the number of buffered data bits in the queue is greater than 20 ms. The queue length PDF of the BIP and iterative schedulers are almost identical but are more spread out than the PDF of the PF scheduler. However, maximum queue lengths are less than half of the maximum allowed delay (20 ms) for the BIP and iterative schedulers, and less than one-tenth for the PF scheduler. Fig. 2.11 shows the average transmission rate for all bearers and schedulers. All the rates converge to the arrival rates and satisfy the rate requirements, which also implies that the data queues are stable. The increase in time complexity of the three schedulers

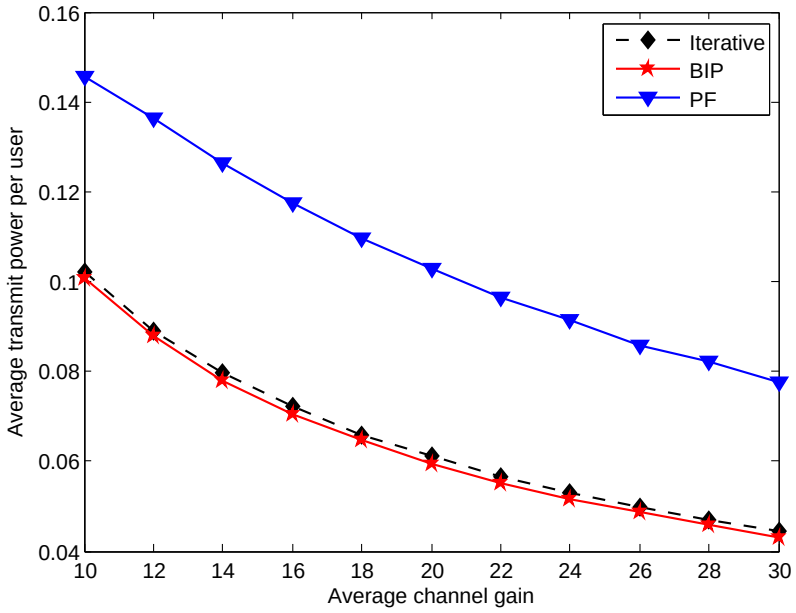

Figure 2.7: Experiment 1: Average transmitted power per user per TTI (Watt).

are compared in Fig. 2.12. The MATLAB functions *tic* and *toc* were used to measure the running time. As expected, the running time of the iterative scheduler is significantly less than that for BIP and PF. Increasing ACG has no effect on the iterative and BIP running times. However, it is directly proportional to the PF time complexity. As ACG increases, more MCSs are considered in each RB chunks, which increases the size of the search space evolved in each iteration.

2.9.2 Experiment 2: Two Users with Identical Conditions but Different ACG

TTo illustrate the behavior of the PF scheduler, two users with different ACGs are considered. The ACG of user one is fixed to 10, whereas the ACG for user two varies from 10 to 30 (10 dB to 14.8 dB). The transmission power for user one and user two are shown in Fig. 2.13 and Fig. 2.14, respectively. Increasing the ACG of user two reduces transmit power consumption. However, this increase has no effect on the power consumption of user one who experiences a constant average channel fading.

Fig. 2.15 shows the queue length in bits for all users and bearers. Successfully, the

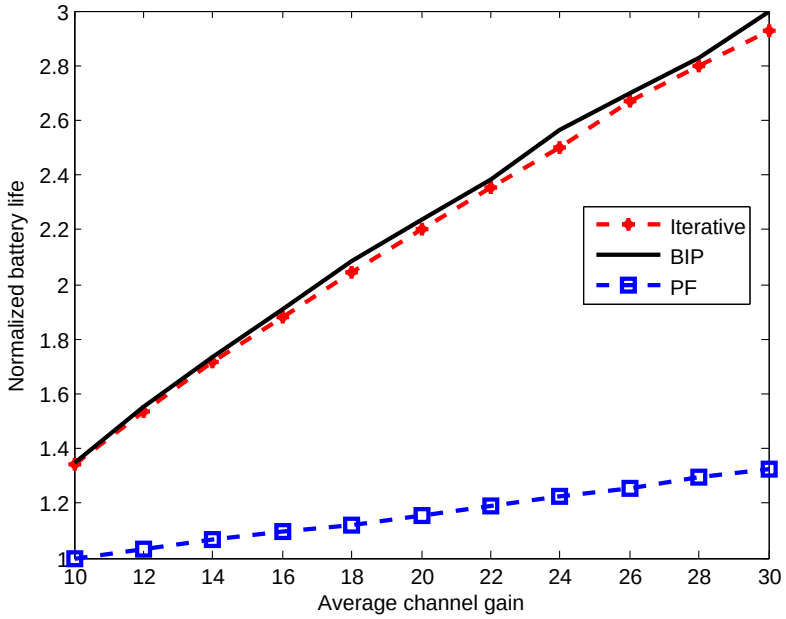

Figure 2.8: Experiment 1: Battery life comparison.

three schedulers maintain average queue lengths of the NGBR bearers less than the maximum allowed average queue length ($\bar{q}_k^n < \hat{q}_k^n$). As user two ACG increases, shorter queue lengths are observed. However, user one queue length remains unchanged. In summary, the proposed schedulers isolate the users from each other. Users who experience good channel conditions consume less power than users who experience severe fading conditions. Fig. 2.16 shows that the transmission rates are equal to the data arrival rates, which implies that all the data arrived has been transmitted.

2.9.3 Experiment 3: The Iterative Algorithm Evaluation

This experiment evaluates the scheduling performance on a relatively large-scale scenario. Due to the fact that the BIP and PF schedulers are computationally heavy, and their time complexity increases exponentially with the number of users as discussed in Sections 2.5.2 and 2.7, simulation experiments are performed only for the iterative scheduler. The proposed algorithm is compared with the EARA scheduler reported in [20]. Note that EARA does not consider the maximum transmit power threshold nor the dynamic traffic behavior. Similar to the proposed algorithm, it is assumed that the EARA allocates $E_k[t]$ bits to UE

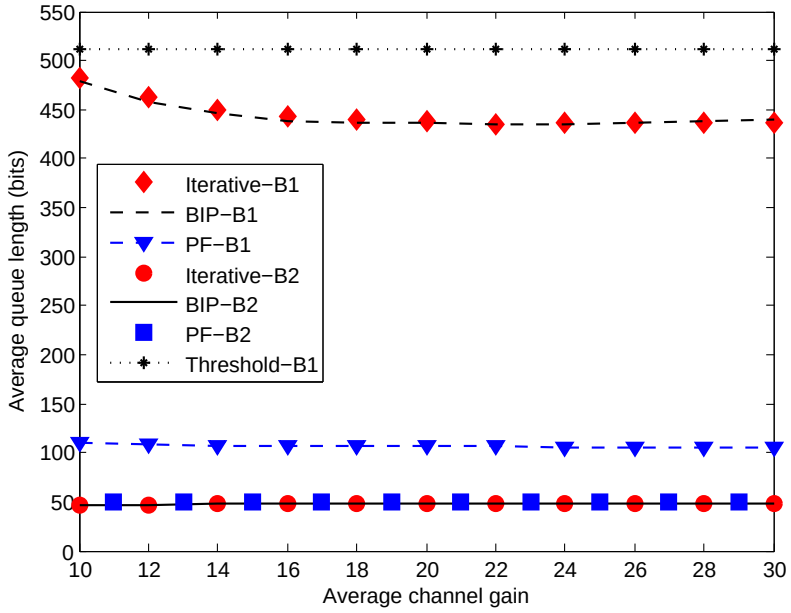

Figure 2.9: Experiment 1: Average queue length per bearer per user.

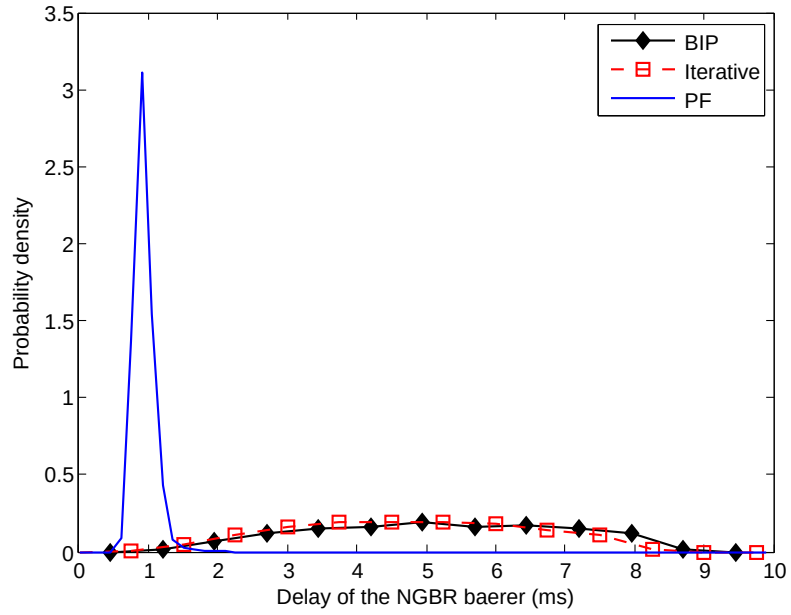

Figure 2.10: Experiment 1: Probability density function of the delay of the NGBR bearers at ACG=10.

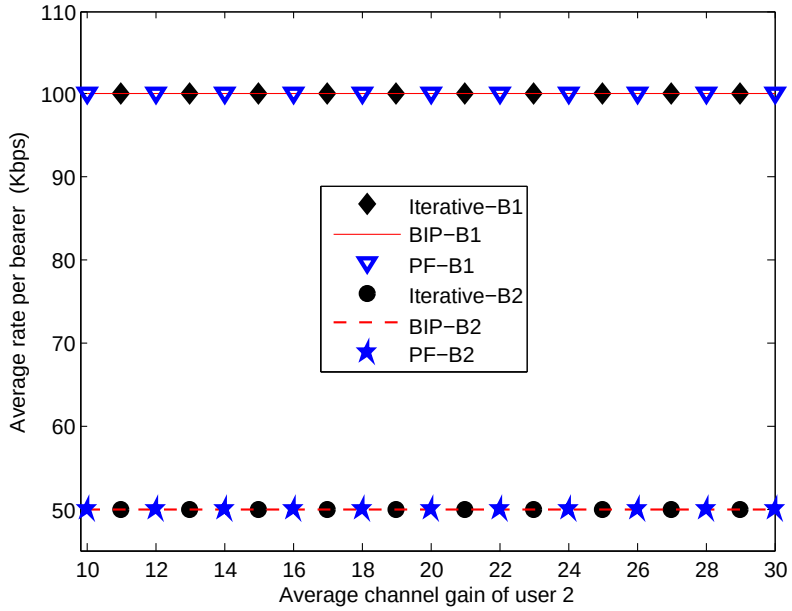

Figure 2.11: Experiment 1: Average transmission rate per bearer per user.

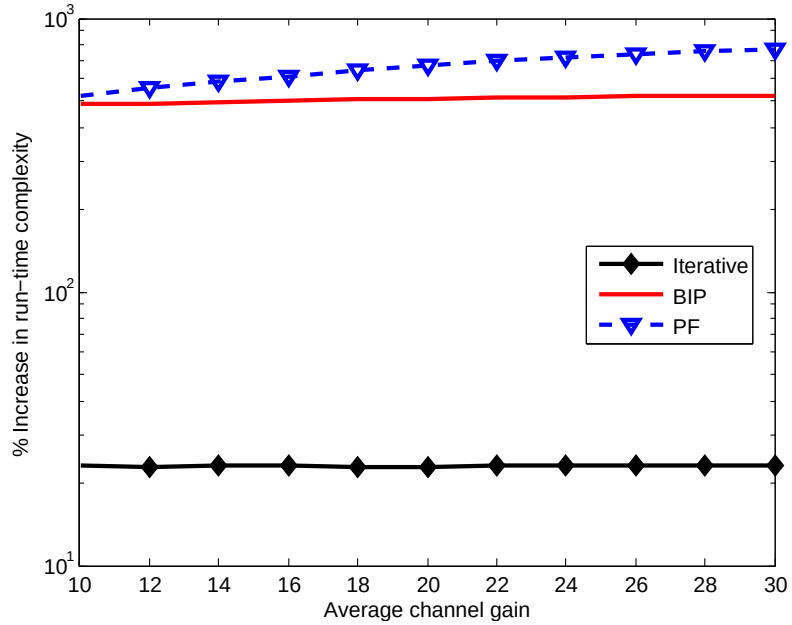

Figure 2.12: Experiment 1: Time complexity comparison.

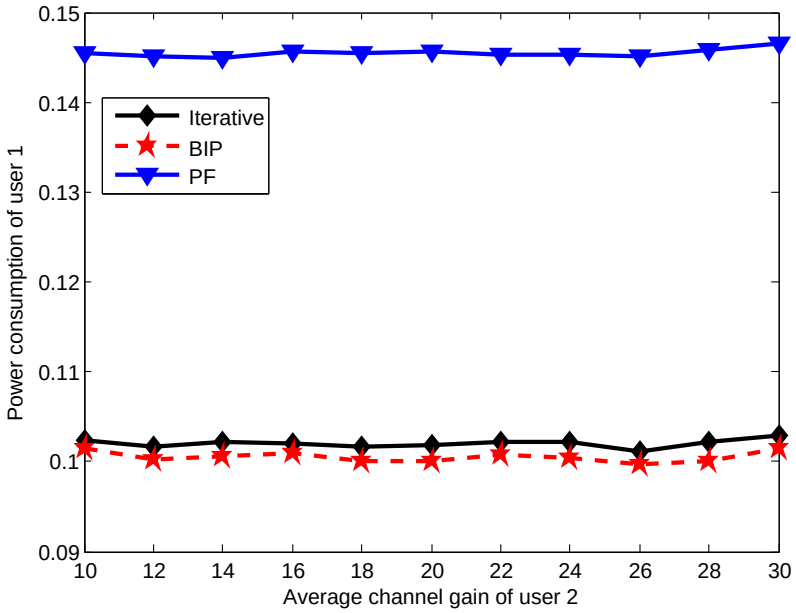

Figure 2.13: Experiment 2: Average transmitted power per TTI for user 1 (Watt).

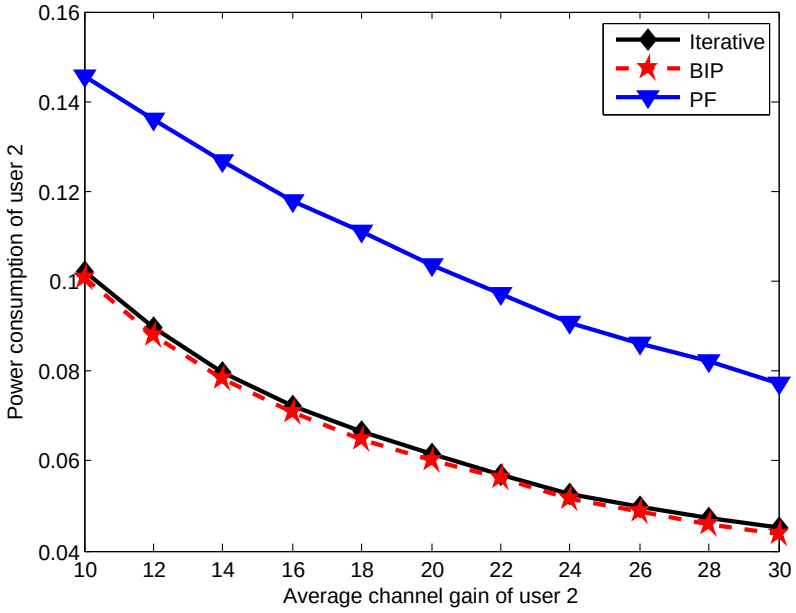

Figure 2.14: Experiment 2: Average transmitted power per TTI for user 2 (Watt).

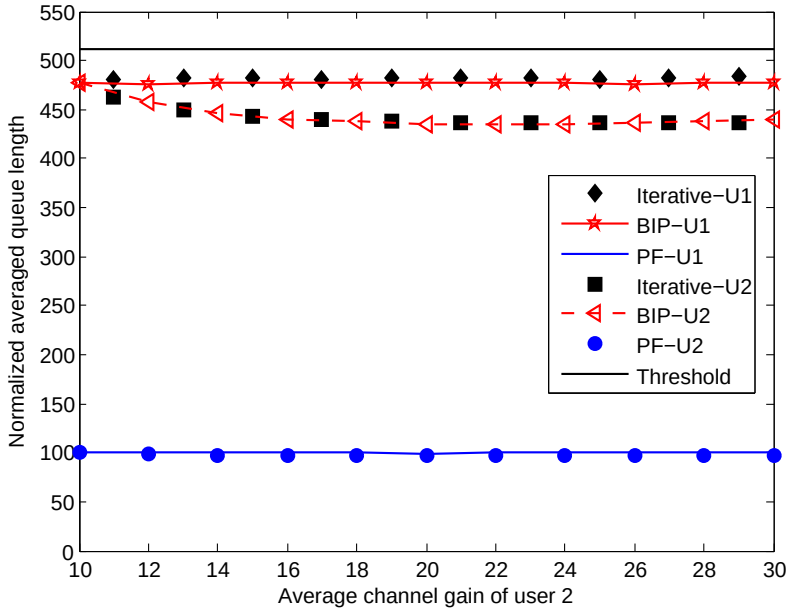

Figure 2.15: Experiment 2: Average queue length for the NGBR bearers in bits.

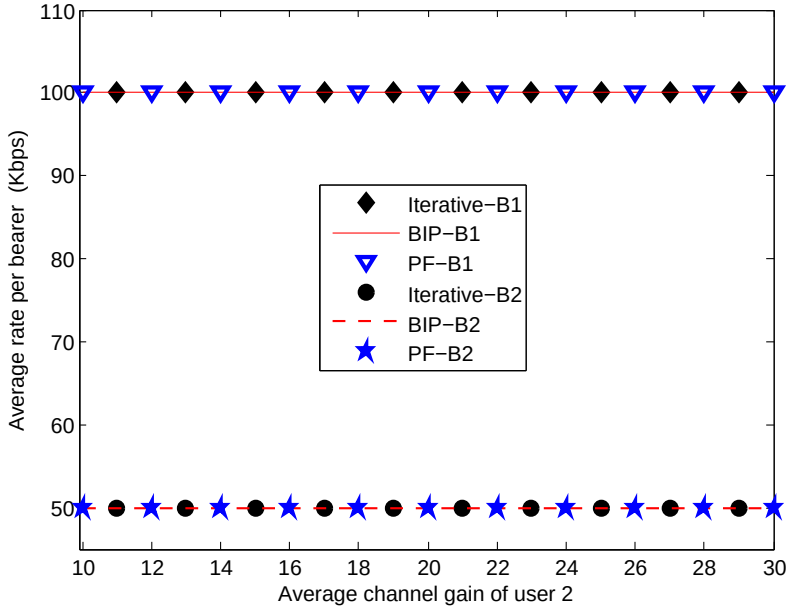

Figure 2.16: Experiment 2: Average transmission rate per bearer per user.

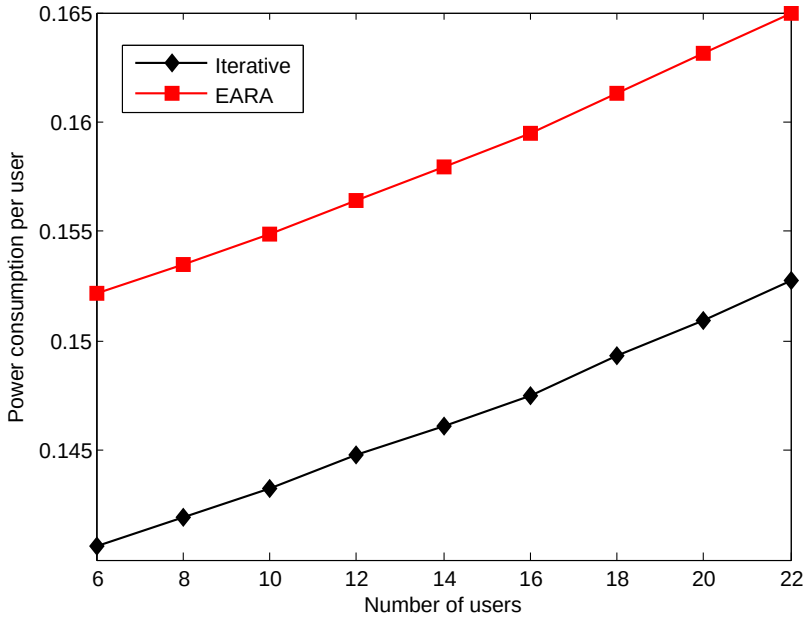

Figure 2.17: Experiment 3: Average transmitted power per user per TTI (Watt).

k at TTI t. The simulation setup is similar to the one given in Table 2.5, but the number of RBs grows to 40.

Each user has four bearers as described in Table 2.6. All users are assumed to experience ACG of 10. Fig. 2.17 exhibits the average transmit power per user. For both algorithms, the number of users is proportional to the average transmission power as a result of the increased competition for resources. Although the EARA is not practical because it does not consider the maximum transmit power threshold, it consumes more power than the proposed algorithm. Fig. 2.18 shows the delay of the four bearers averaged over all users. The delay of the EARA is slightly lower than the delay of the proposed algorithm. As the number of users increases, the delay becomes longer and approaches their thresholds. However, both algorithms succeed to maintain the delay less than the threshold. The average data rates are shown in Fig. 2.19. The algorithms manage to transmit average data rates equal to the average data arrival rates.

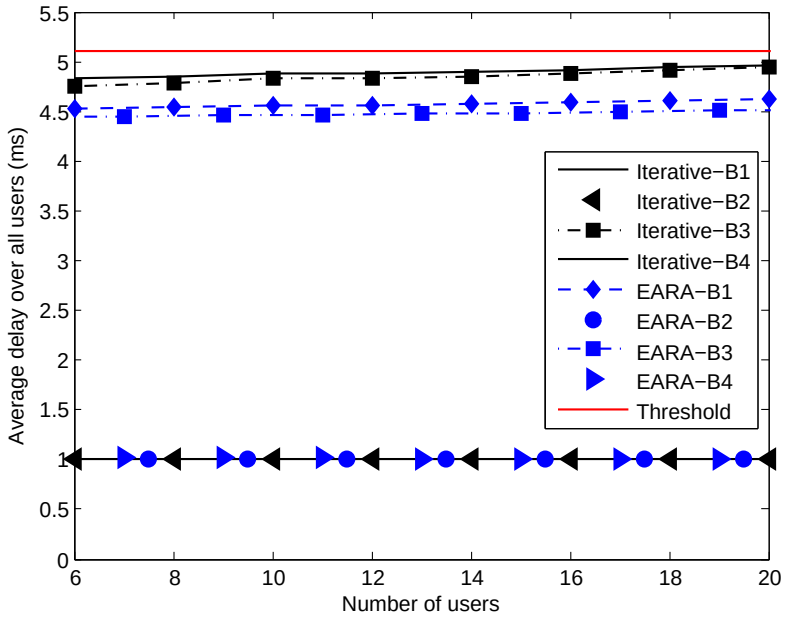

Figure 2.18: Experiment 3: Delay per bearer averaged on all users.

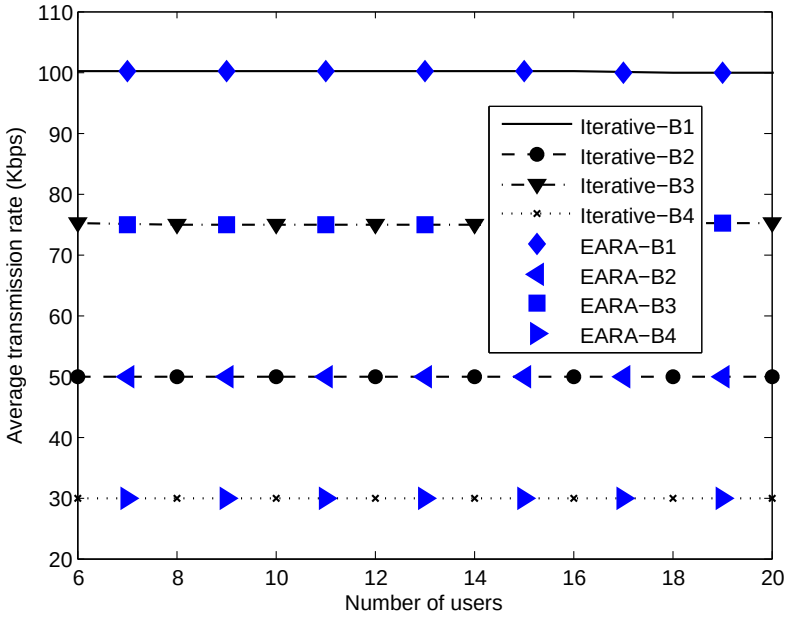

Figure 2.19: Experiment 3: Average transmission rate per bearer averaged on all users.

2.10 Chapter Summary

In this chapter, a framework for power-efficient scheduling in LTE uplink systems is presented. Both the QoS requirements and the channel fading parameters were considered. The scheduling problem was formulated and presented as a multi-stage problem. Then, it was simplified into a single point binary integer programming problem. Subsequently, a low-complexity iterative scheduler was proposed to solve the binary integer programming problem. The iterative scheduler proved to consume slightly more power compared to the binary integer programming scheduling approach, but it has considerably lower computational complexity. Simulation results were used to compare the proposed schedulers with the Proportional Fair scheduler in terms of power efficiency, delay, transmission rate, and complexity. The results show that the proposed schedulers maintained the required QoS and reduced the total transmit power under different practical scenarios. These power savings were achieved because of the schedulers' attribute of transmitting data at low rates while maintaining the required QoS.

Chapter 3
Low-Complexity Power-Efficient Schedulers for LTE Uplink with Delay-Sensitive Traffic

3.1 Introduction

The recent years have witnessed unremitting advances in the wireless technology domain, which served to grow the mobile data market. New wireless applications and services have emerged, and accessing data services via mobile devices has increased considerably. To keep up with the increase in mobile data traffic, long term evolution (LTE) technology has been developed to support high performance radio-access technology.

LTE supports high data rate links and enables users to run multiple concurrent applications with heterogeneous quality of service (QoS) requirements, such as live streaming of audio, video, and social media applications. However, to maintain a fixed error performance, increasing the transmitted data rate is accompanied with a power increase to keep the energy per bit unchanged. Furthermore, as the total number of bits transmitted per unit time grows, the total transmission power per unit time becomes substantially higher as well. Unfortunately, the increasing demand for transmission power is quite higher than the improvement in batteries' capacity. As most end-user devices are powered from small size batteries, high data rate transmission would reduce the average operation time-per-charge of battery-powered devices. Consequently, the development of power-efficient transmission techniques has become an important design consideration to improve the battery life of mobile devices.

A version of this chapter has been published in [34].

In the literature, there has been increasing interest to better understand and model the power consumption of smartphones. For example, Zhangt *et al.* [35] designed an online power model that estimates the power consumption of different components in Android smartphones including central processing unit (CPU), liquid-crystal display (LCD), global positioning system (GPS), audio, Wi-Fi and cellular interfaces. The work reported in [36] models the impact of wireless signal strength on smartphone energy by analysing traces collected from 3785 smartphones. A power model of a commercial LTE network is presented in [37], where an application is designed and installed on Android smartphones to collect traces of the power consumption. The study suggests that the power consumption of LTE is 23 times higher than the power consumption of WiFi interfaces.

In LTE systems, the LTE uplink is based on single carrier frequency division multiple access (SC-FDMA). Compared to orthogonal frequency division multiple access (OFDMA), SC-FDMA has lower peak-to-average power ratio (PAPR). The Low PAPR advantage of SC-FDMA is achieved by localized-mapping of the resource blocks (RBs), where each user should be mapped to a subset of contiguous RBs.

Resource scheduling in OFDMA-based systems has been widely investigated in the literature [38]. Many schedulers have been developed to optimize different allocation metrics such as the sum rate maximization [39], total transmit power minimization [40], and fairness [41]. Several solutions have been presented based on game theory [41], convex optimization and dual decomposition [42, 43, 44], dynamic backpressure policies [45], and interior point methods [46]. However, the contiguity constraint of the SC-FDMA changes the scheduling problem into a non-convex optimization problem [13, 47], and prevents the direct application of power-efficient transmission techniques that are derived for OFDMA systems [38]. Due to the contiguity constraint, the resource allocation in SC-FDMA systems is typically formulated as a binary integer programming (BIP) problem [8, 13, 20, 25, 47].

In the literature, the resource allocation for SC-FDMA systems can be divided into two groups. The first group considers a static data traffic model. The main objective in such scenarios is either to maximize the aggregated cell throughput subject to a maximum transmit power threshold, or to minimize the total transmit power for all users subject to a constant data rate. For example, Wong *et al.* [47] considered weighted sum-rate max-

imization in the LTE uplink. The problem is formulated and solved as a BIP. A reduced complexity technique that solves the same problem is reported in [13], where the BIP is transformed into a continuous space canonical dual problem, which is solved using algorithms with polynomial complexity. Heuristic algorithms were also proposed to solve the BIP problem with lower complexity [48, 49, 50].

It is worth noting that the schemes described in [13, 47, 48, 49, 50] aim at maximizing the system capacity regardless of the power consumption of user equipments (UEs). The schemes assume that UEs transmit at their maximum transmit power, which degrades the power efficiency of the scheduler. Moreover, the schemes do not consider the dynamic nature of the traffic and assumes full-buffer occupancy, which is not necessarily true in practical systems.

The second group considers dynamic traffic models and QoS requirements as the optimal scheduler presented in Chapter 2. However, the proposed scheduler is complex and not globally optimal. In addition, it requires that the average arrival rates of all user traffic bearers to be known at the evolved node-B (eNB).

In this chapter, the global optimal scheduler for the LTE uplink is derived. The scheduler minimizes the total transmit power of all users while satisfying the delay requirements. The scheduling problem is formulated as a dynamic programming (DP) problem, and the scheduler considers the dynamic nature of the traffic load, maximum transmit power threshold, contiguous allocation, and the time-varying fading channel. Moreover, to reduce the complexity, two power-efficient heuristic schedulers are proposed to solve the scheduling problem.

The rest of this chapter is organized as follows. Section 3.2 presents the system model. Section 3.3 discusses the scheduling constraints. The DP problem is formulated and discussed in Section 3.4. The heuristic algorithms are described in Section 3.5. The numerical results are presented in Section 3.6, and finally Section 3.7 concludes the chapter.

3.2 System Model Description and Assumptions

An uplink SC-FDMA multiuser system is considered, where K UEs communicate with an eNB. The uplink bandwidth is divided into M RBs. Each RB consists of 12 adjacent

Table 3.1: Summary of the most significant notation used in this chapter.

Symbol	Meaning
K, k	Number of UEs / UE Index
M, m	Number of RBs / RB index
N, n	Number of subcarriers / Subcarrier index
t	Subframe index
$p_{k,n}$	Power allocated to subcarrier n of user k
P_k	Total transmit power of user k
P^{max}	Transmit power threshold
$\alpha_{k,m}$	Channel-gain-to-noise ratio of user k on the RB m
q_k	The queue length of user k (DUs)
q_k^{max}	Queue length threshold of user k (DUs)
T_k	Transport block size (bits per subframe)
L_k	User k DU length (in bits)
c_k	Transmitted DUs from buffer k
c_k^{max}	Max. number of DUs/subframe user k can transmit
b_i	The contiguous allocation number i
b	Set of all contiguous RB allocations
c	Joint actions of all users
s	Joint state of all users
β	Binary number indicator
a	All users' arrival DUs

subcarriers. Therefore, the number of subcarriers available for the uplink transmission is $N = 12\ M$. The available RBs are assigned to users by the eNB. Each user can be assigned one or more RBs. To facilitate the readability, Table 3.1 summarizes the notations frequently used throughout the chapter.

The SC-FDMA transmission process is shown in Fig. 3.1. Without loss of generality, suppose that user k is allocated RBs $\{1, 2, ..., M_k\}$. In other words, the subcarriers $\{1, 2, ..., N_k\}$ are allocated to user k, where $N_k = 12\ M_k$. In this section, the temporal index t_s is omitted for brevity. The transmitted signal of user k, without cyclic prefix, can be expressed as

$$\mathbf{s}_k = \mathbf{G}\mathbf{Z}_k\mathbf{F}_k\mathbf{P}_k\mathbf{x}_k \tag{3.1}$$

where $\mathbf{G} \in \mathbb{C}^{N\times N}$ is the inverse Discrete Fourier Transform (IDFT) matrix, $\mathbf{Z}_k \in \mathbb{R}^{N\times N_k}$ represents the mapping matrix for subcarrier assignment, $\mathbf{F}_k \in \mathbb{C}^{N_k\times N_k}$ is Discrete Fourier Transform (DFT) matrix, $\mathbf{P}_k$ is diagonal power matrix $\mathbf{P}_k = diag\{\sqrt{p_{k,n}} : n = 1, 2, ..., N_k\}$,

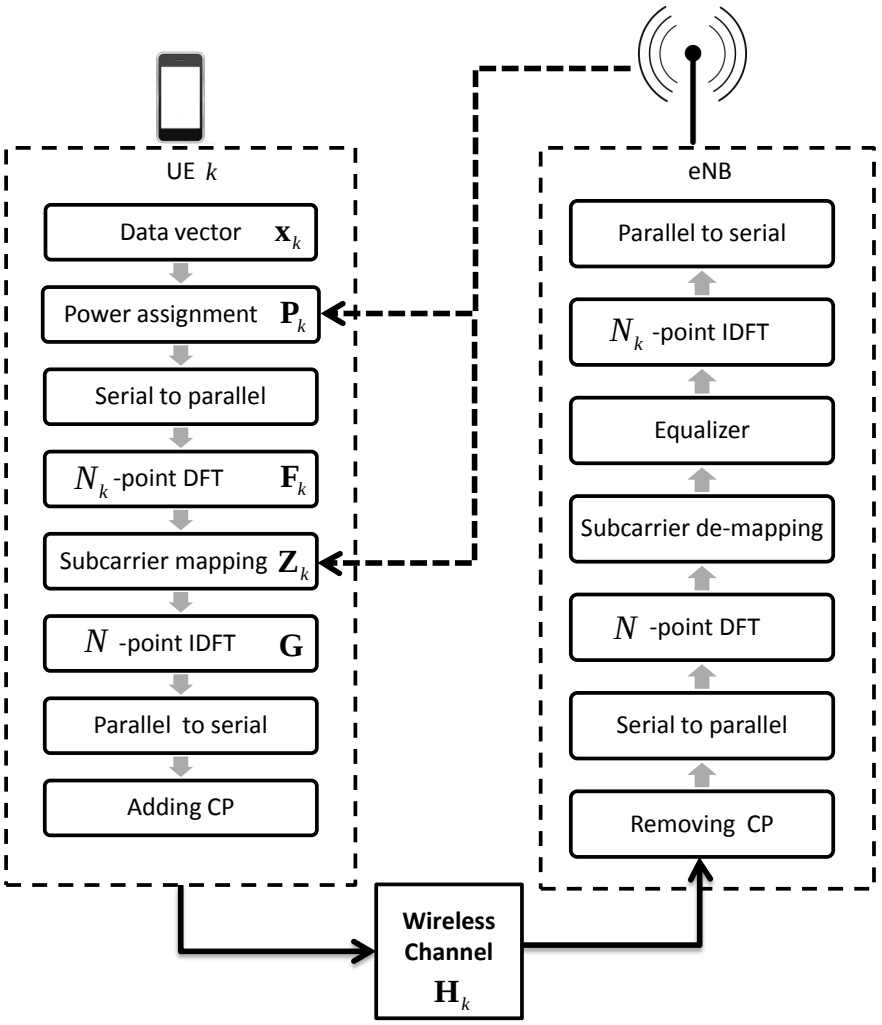

Figure 3.1: A block diagram of a SC-FDMA system.

with $p_{k,n}$ represents the power allocation to the subcarrier n, and $\mathbf{x}_k$ is the $N_k \times 1$ data vector, which represents the time-domain symbols sent by user k. It is worth noting that the mapping matrix $\mathbf{Z}_k$ changes the size of vector $\mathbf{F}_k\mathbf{P}_k\mathbf{x}_k$ from $N_k \times 1$ to $N \times 1$. The mapping

matrix of user k is defined as follows

$$\mathbf{Z}_k = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{3.2}$$

To reduce the signalling overhead, the LTE standard specifies that when a user is assigned more than one RB, one power level should be used for all RBs [9, 15, 51]. Therefore, the transmit power is equally divided over all subcarriers

$$p_{k,n} = \frac{P_k}{N_k}, \quad n = 1, 2, ..., N_k \tag{3.3}$$

where P_k is the total transmit power of user k. Consequently, the power matrix can be expressed as

$$\mathbf{P}_k = \frac{P_k}{N_k}\mathbf{I} \tag{3.4}$$

where $\mathbf{I} \in \mathbb{R}^{N \times N}$ is an identity matrix. The cyclic prefix length is assumed to be longer than the maximum excess delay of the channel. Therefore, the inter-symbol interference can be eliminated by removing the cyclic prefix.

In LTE systems, fractional frequency reuse (FFR) techniques are used to control the intercell interference of the cell-edge users by exchanging interference-coordination information between eNBs over the X2 interface. With FFR, users are assigned different parts of wireless resources based on their location in the cell. Users who are close to the center of the cell are subject to low intercell interference from neighbouring cells and are allowed to use the entire frequency band (frequency reuse factor of one). However, users at cell-edges are subject to high intercell interference. Exclusive frequency bands, that are not used in neighbouring cells, are assigned to cell-edge users to reduce the intercell interference. In this chapter, it is assumed that the scheduling is performed independently

at each base station because the interference at the edges is generally limited due to FFR techniques [52].

At the eNB, the received signal for all users, after removing the cyclic prefix and performing the N-point DFT, can be expressed as

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{H}_k \mathbf{Z}_k \mathbf{F}_k \mathbf{P}_k \mathbf{x}_k + \mathbf{w}. \tag{3.5}$$

Assuming that the channel remains unchanged over one SC-FDMA subframe, then the channel matrix $\mathbf{H}_k \in \mathbb{C}^{N \times N} = diag\{H_{k,n} : n = 1, 2, ..., N\}$ is the diagonal channel matrix between user k and the eNB, $H_{k,n}$ is the channel gain of the n^{th} subcarrier seen by user k includes path loss, shadowing, and multipath fading, and $\mathbf{w}$ is the additive white Gaussian noise vector with variance N_o.

In LTE, the eNB manages the resource allocation in both downlink and uplink [9]. The allocation of subcarrier and power is determined for each user by the eNB based on the scheduling policy. In other words, the eNB determines the matrices $\mathbf{Z}_k$ and $\mathbf{P}_k$ for each user. Users are assumed to experience independent Rayleigh block-fading channels. Each channel is assumed to be fixed during one subframe in time domain and over one RB in frequency domain $\Big(H_{k,12(m-1)+v} = H_{k,12(m-1)+u}, [v, u] \in \{1 : 12\}\Big)$, but changes independently over different subframes and different RBs. Assuming that the eNB applies the zero-forcing equalizer, the post-processing time-varying effective SNR for the received signal of user k can be expressed as [53]

$$\gamma_k = P_k \left(\sum_{m=1}^{M_k} \frac{1}{\alpha_{k,m}} \right)^{-1} \tag{3.6}$$

where $\alpha_{k,m}$ is the channel-gain-to-noise ratio (CNR) of user k on the RB M, which is defined as

$$\alpha_{k,m} = \frac{|H_{k,12(m-1)+v}|^2}{N_o}, v \in \{1 : 12\}. \tag{3.7}$$

The granularity of the scheduler decisions is one RB in the frequency domain and one LTE subframe in the time domain, which is also called a transmission time interval (TTI). Fig. 3.2 illustrates the LTE subframe structure. Each LTE subframe contains 14 symbols,

each of which has a duration of $T_s = 66.67\mu s$. Suppose that T_{ctrl} symbols are used for signalling, the number of data symbols per subframe over an RB is equal to $12\,(14 - T_{ctrl})$. The total number of bits per subframe that user k can transmit is

$$T_k = R_k\, 12\, (14 - T_{ctrl})\, T_s \tag{3.8}$$

where R_k is the maximum instantaneous data rate (bits/second) that user k can achieve using the assigned RBs $\{1, 2, ..., M_k\}$, which can be calculated as [29, 54]

$$R_k = W_s\, N_k \log_2 (1 + \gamma_k) \tag{3.9}$$

where W_s is the subcarrier bandwidth. Using (3.6), (3.8), and (3.9) the required power to transmit T_k bits for user k over the assigned subcarriers $\{1, 2, ..., M_k\}$ is

$$P_k = (2^{\kappa} - 1) \sum_{m=1}^{M_k} \frac{1}{\alpha_{k,m}} \tag{3.10}$$

where

$$\kappa = \frac{T_k}{12\, (14 - T_{ctrl})\, T_s\, W_s N_k}. \tag{3.11}$$

Uplink synchronization is needed to avoid overlapping transmissions from different users. In LTE systems, the uplink transmission timing between different users is achieved by using the timing advance mechanism [9]. The eNB keeps measuring the timing of the user uplink signal, and sends timing advance commands to the user who needs a timing adjustment. The timing advance commands update the uplink transmission timing and offset the differing propagation delays of users. In this chapter, it is assumed that the eNB provides perfect timing and frequency synchronization [55].

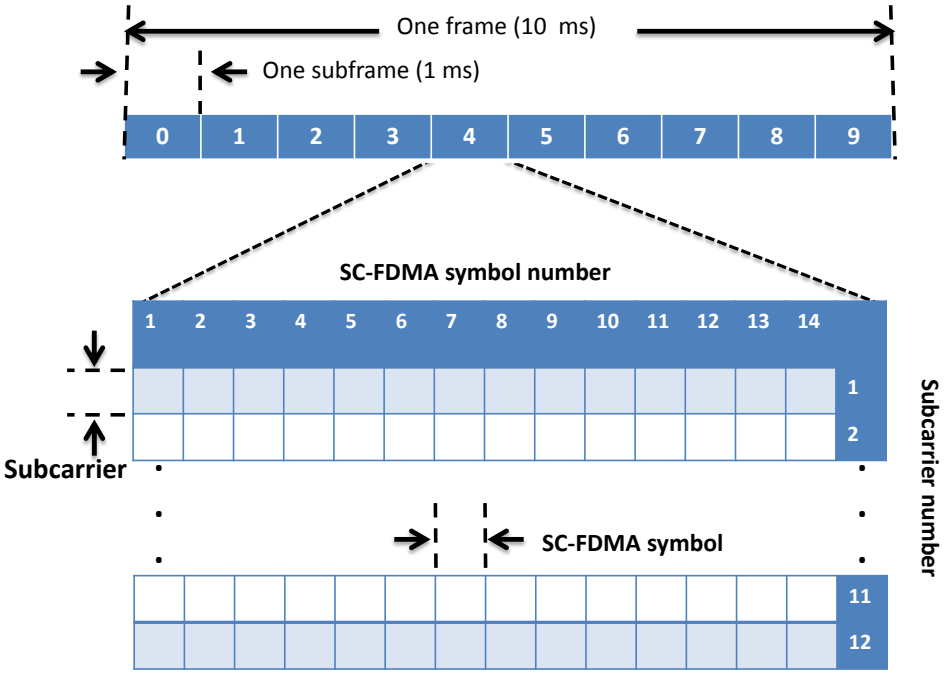

Figure 3.2: The structure of the LTE frame.

3.3 Allocation Constraints and Problem Definition

3.3.1 Delay Constraint

DUs arrive randomly to users' buffers where each user is assumed to have one finite-length first-input-first-output buffer to store the unserved DUs.

Each buffer is assumed to be large enough to accommodate all unserved DUs.

Let $a_k[t]$ denote the number of DUs that arrive to the user buffer k at the end of subframe t. The distribution of the DUs arrival follows Poisson distribution with associated parameter λ_k [30]. User k buffer updates its status over any two successive subframes as

$$q_k[t+1] = q_k[t] + a_k[t] - c_k[t] \tag{3.12}$$

where $q_k[t]$ denotes the number of DUs in buffer k at the beginning of subframe t, and $c_k[t]$ is the transmitted DUs from buffer k during $[t, t+1)$.

The LTE standard defines a packet delay budget (PDB) for each bearer, which defines the maximum acceptable time that a packet may be delayed between the UE and the PCEF

(Policy and Charging Enforcement Function) [56]. For example, conversational voice has PDB of 100 ms. However, the buffer status reporting (BSR) mechanism in LTE informs the eNB the data size (in bytes) in the user buffer [9]. In the downlink, the delay of each packet is known by the eNB. In the uplink, BSR mechanism is responsible for updating the eNB the current status of users' buffers. However, the BSR reports contain only the queue length information, which is used to perform the scheduling.

Network operators can determine the upper bound on the user delay based on statistical measurements of the queue length, if the queue length exceeds the upper bound, delay violation is declared.

In this chapter, it is assumed that the delay requirements for the data delivery of user k is satisfied if

$$q_k[t] \leq q_k^{max}, \forall t \tag{3.13}$$

where q_k^{max} is an integer number associated with the PDB. However, once the number of DUs exceeds q_k^{max}, the delay violation is declared, but no DU is discarded.

3.3.2 System Constraints

Three types of constraints restrict the scheduling in the LTE uplink: 1) Contiguity constraint: SC-FDMA restricts the RB allocation to only contiguous ones, meaning that if a user is assigned more than one RB, the assigned RBs should be adjacent to each other. The set that contains all possible contiguous allocations is denoted as $\mathbf{b} = \{b_1, b_2, ..., b_{tot}\}$, where b_i is the contiguous allocation number i, and b_{tot} is the total number of possible contiguous allocations that can be constructed from x RBs, which can be calculated as [47]

$$b_{tot} = \frac{1}{2}\, x(x+1) + 1. \tag{3.14}$$

Fig. 3.3 shows an example of the set **b** for three RBs. The contiguity constraint can be maintained by allocating adjacent RBs to every user. 2) Exclusivity constraint where each RB should not be assigned to more than one user. 3) Power constraint that limits the the maximum transmit power per user to a threshold specified by the LTE standard. This restriction is required to ensure that the transmitter power amplifier is operating in the linear

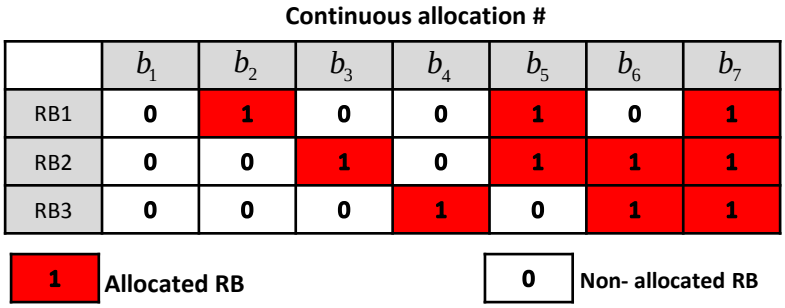

Figure 3.3: Contiguous allocations possible from three RBs.

region. The power constraint is satisfied by maintaining the following inequality

$$P_k[t] \leq P^{max}, \forall t \tag{3.15}$$

where P^{max} is the transmit power threshold.

3.3.3 Problem Definition

This chapter aims at minimizing the total power transmitted by all users while satisfying the delay constraints. This optimization problem can be formulated as

$$\min \lim_{\tau \to \infty} \frac{1}{\tau} \sum_{k=1}^{K} \sum_{t=1}^{\tau} P_k[t] \tag{3.16a}$$

subject to:

$$q_k[t] \leq q_k^{max}, \forall k \tag{3.16b}$$

$$I_k[t] \in \mathbf{b}, \forall k \tag{3.16c}$$

$$I_u[t] \cap I_v[t] = \phi, \forall u \neq v \tag{3.16d}$$

$$P_k[t] \leq P^{max}, \forall k \tag{3.16e}$$

where $I_k[t]$ represents the contiguous allocation assigned to user k at subframe t. The constraints in (3.16b-3.16e) assure that the delay, exclusive allocation, contiguity allocation, and maximum allowable power transmit constraints are satisfied.

In this chapter, it is assumed that the system can maintain the delay requirements of all users, and feasible solutions exist. However, if the data arrival rates are high, the optimization in (3.16) could be infeasible, and hence, an admission control policy is needed [12].

3.4 Optimal Offline Scheduling

In this section, an offline optimal solution for (3.16) is obtained using DP. The problem in (3.16) is optimally structured, i.e., it is divided into stages each of which has a duration of one subframe. By assuming that the scheduler has perfect knowledge of both future DU arrivals and future channel realizations before scheduling, the problem in (3.16) can be expressed as a deterministic DP problem. Although this assumption makes the system non-causal, it provides the optimal solution which can be used to as a reference for other sub-optimal algorithms.

Having said that, it is worth noting that such assumptions are not always over optimistic because DU arrivals can be predicted for certain applications such as voice over IP (VoIP), and future channel state information can be accurately estimated in slow time varying channels [57, 58].

In this chapter, it is assumed that the system is observed over a finite number of subframes (stages) τ. Three main elements define the DP, namely stages, states, and actions, which are described as follows.

- **Stages**: DP breaks up the entire problem into stages, and each stage forms a new sub-problem. The new sub-problems are smaller, and consequently, less computationally-expensive to solve compared to the entire problem. The optimal solution of the entire problem can be achieved by solving all sub-problems individually. For the problem considered in this chapter, the stage is defined as one subframe in time.
- **States**: every stage has a number of possible states. Information about the states at each stage is essential to solve the sub-problems. To avoid confusion, two types of states, user states and joint states are defined. At subframe t, the state of user k is denoted by $q_k[t]$ and defined as the number of DUs that exist in the buffer of user k. The joint state at stage t is defined as the joint states of all users at subframe t, which

can be expressed as

$$\mathbf{s}[t] = [q_1[t], q_2[t], ..., q_K[t]]. \tag{3.17}$$

Given that the the maximum allowable delay of user k is q_k^{max} DUs, the possible states of user k are $[0, 1, ..., q_k^{max}]$. User k has $q_k[t] + 1$ possible states at stage t. Therefore, the total number of joint states at stage t is

$$N^{js}[t] = \prod_{k=1}^{K} (1 + q_k[t]). \tag{3.18}$$

- **Actions**: are defined as the number of DUs a user transmits at a stage. Actions are taken at every stage for all users and update their states. As users compete for the same radio resources, the actions for each user depend on the action taken for the other users. As a result, the joint actions should be taken into account by the dynamic algorithm. The joint actions are defined as number of transmit DUs per subframe for each user. Fig. 3.5 illustrates the components and the optimal action path of the DP problem.

  To simplify the DP analysis, it is assumed that users can only send a finite number of DUs during any TTI. The action set that user k can take at stage t is denoted by

$$c_k[t] \in \{0, 1, 2, ..., \min(q_k[t], c_k^{max})\}, \forall k \tag{3.19}$$

  where c_k^{max} is the maximum number of DUs that user k is allowed to transmit per subframe. Note that (3.19) implies that user k cannot take any action at stage t that transmits data more than its buffer length.

  The joint action at stage t is described as joint transmission decisions for all users, and it is denoted by

$$\mathbf{c}[t] = [c_1[t], c_2[t], ..., c_K[t]]. \tag{3.20}$$

  Therefore, the number of joint actions at stage t is

$$N^{ja}[t] = \prod_{k=1}^{K} (1 + min(q_k[t], c_k^{max})). \tag{3.21}$$

In fact, different possible distributions of the RBs between users may exist for one joint action. Consequently, different RB allocations may result in different costs with the same joint action. For example, assume a system of two users who share three RBs. All possible RB allocations for the two users for each action are shown in Fig. 3.4. As channel gain varies over RBs, each RB allocation is associated with a cost which can be calculated using (3.10). As a result, each joint action in this example may be related to nine different joint costs depending on how the RBs are assigned to the users.

As our objective is to minimize the total transmit power, finding the minimum joint cost for each action results in minimizing the total transmit power. Therefore, for each action, the RBs should be allocated to the users such that the cost is minimized. Nevertheless, finding the actions' minimum joint costs is not trivial, particularly for systems with large number of users and RBs.

Joint Action Cost Minimization: finding the minimum cost for each joint action is a combinatorial problem, which can be formulated as a binary integer programming problem as follows. Assume the current subframe is t, the minimum cost for the joint action $\mathbf{c}[t] = [c_1[t], c_2[t], ..., c_K[t]]$ can be found by solving the following optimization problem

$$P(\mathbf{c}, t) = \min \sum_{k=1}^{K} \sum_{b=1}^{B} P_k[t]\, \beta(k, b_i) \tag{3.22a}$$

subject to:

$$\bigcap_{k=1}^{K} \beta(k, b_i) = \phi,\ \forall i \tag{3.22b}$$

$$P_k[t] \leq P^{max} \tag{3.22c}$$

$$T_k[t] = c_k[t]\, L_k,\ \forall k \tag{3.22d}$$

where L_k is user k DU length (in bits), $\mathbf{c}$ is a generic value of $\mathbf{c}[t]$, and $\beta(k, b_i)$ is a binary number defined as

$$\beta(k, b_i) = \begin{cases} 1, \text{ if the allocation } b_i \text{ is assigned to user } k. \\ 0, \text{ otherwise.} \end{cases} \tag{3.23}$$

Although data is delivered in packets of several sizes, they may be split into several DUs or may be combined into a single DU. Here, it is assumed that L_k is a fixed value for each user; but a user can transmit one or more DUs in each TTI. For example, if $L_k = 100$ bits and $c_k^{max} = 3$, user k can transmit 100, 200, or 300 bits in each TTI.

The constraint in (3.22b) maintains exclusive RB allocations, and prevents overlapping allocations. It is worth noting that feasible solutions for (3.22) are not guaranteed. If no feasible solutions exists, the cost of the joint action is assigned as ∞.

A main limitation for the DP is that the complexity of such algorithms increases exponentially as a function of the number of possible actions at each stage. Although solving BIP for each joint action is computationally expensive, it results in reducing the computational complexity of the DP algorithm. Instead of considering all RB allocation possibilities for one joint action as new joint actions, they reduce to only one joint action.

The Basic DP Algorithm: the joint state updates as follows

$$\mathbf{s}[t+1] = \mathbf{s}[t] - \mathbf{c}[t] + \mathbf{a}[t], t \geq 1 \tag{3.24}$$

where $\mathbf{s}[1]$ is the initial state and $\mathbf{a}[t]$ presents all users' arrival DUs at subframe t

$$\mathbf{a}[t] = [a_1[t], a_2[t], ..., a_K[t]]. \tag{3.25}$$

The problem in (3.16) can be presented as an optimal control problem as follows

$$\min_{\mathbf{c}} \frac{1}{\tau} \sum_{t=1}^{\tau} P(\mathbf{c}, \mathbf{s}, t) \tag{3.26}$$

where $P(\mathbf{c}, \mathbf{s}, t) = P(\mathbf{c}, t) + P_I(\mathbf{s}, t)$, and $P_I(\mathbf{s}, t)$ is a penalty cost of infeasible states which is defined as

$$P_I(\mathbf{s}, t) = \begin{cases} \infty, \text{ if } \mathbf{s} \text{ is infeasible} \\ 0, \text{ if } \mathbf{s} \text{ is feasible.} \end{cases} \tag{3.27}$$

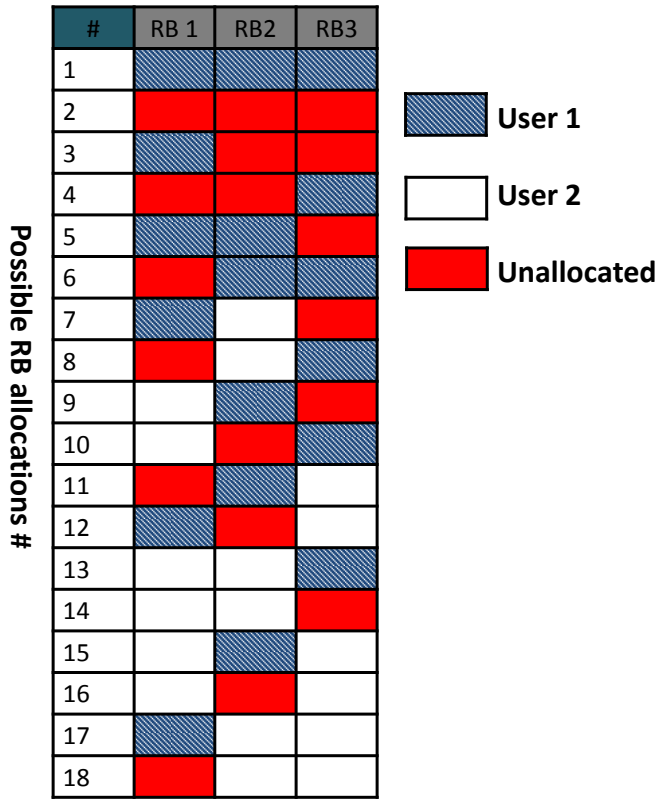

Figure 3.4: An example of all possible RB allocations for a system consisting of three RBs shared between two users.

User k reaches an infeasible state at subframe t if the number of the unserved bits in the user's buffer exceeds the maximum allowed queue length, $q_k[t] > q_k^{max}$.

Initial and final state settings: assume that all users' buffers are empty at both the initial subframe ($t = 1$) and final subframe ($t = \tau$). Also, no data arrives for any users at time $t \geq \tau$ is assumed. In other words, $\mathbf{a}[t] = \mathbf{0}_K | t \geq \tau$, where $\mathbf{0}_K$ is a vector of zeros with a length of K. Therefore, the first and the $\tau + 1$ joint states can be presented, respectively, as $\mathbf{s}[1] = \mathbf{0}_K$ and $\mathbf{s}[\tau + 1] = \mathbf{0}_K$.

Let $F(\mathbf{s}, t)$ be the minimum future cost obtained by optimising the problem over subframes $t, t + 1, ..., \tau + 1$, where $\mathbf{s}$ is a generic value of $\mathbf{s}[t]$. The dynamic programming

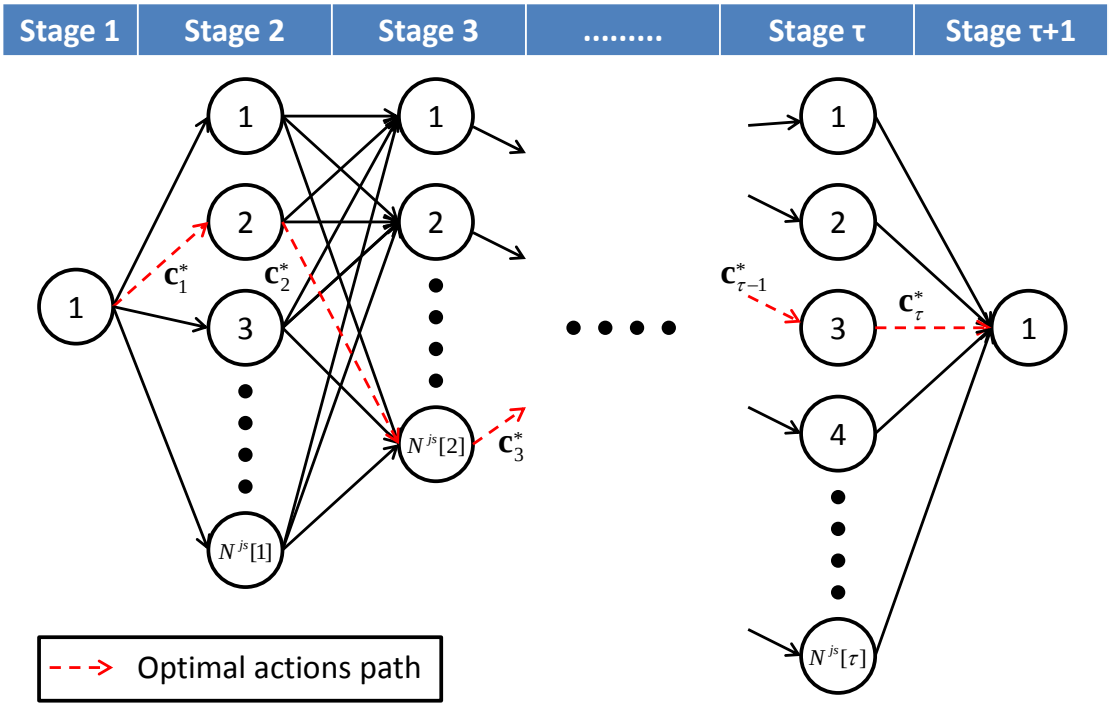

Figure 3.5: The components and the optimal action path of the DP problem.

equation, also known as Bellman equation, can be written as

$$F(\mathbf{s},t) = \begin{cases} \min\limits_{\mathbf{c}} \{P(\mathbf{c},\mathbf{s},t) + F(\mathbf{s}-\mathbf{c}+\mathbf{a},t+1)\}, & t \leq \tau \\ 0, & t > \tau. \end{cases} \tag{3.28}$$

Note that the optimization in (3.28) is unconstrained because all the constraints are considered in $P(\mathbf{c},\mathbf{s},t) = P(\mathbf{c},t)+P_I(\mathbf{s},t)$. The exclusive allocation, contiguity allocation, and maximum allowable transmit power constraints are considered in $P(\mathbf{c},t)$ as illustrated in (3.16). The delay constraint is considered in $P_I(\mathbf{s},t)$ as discussed in (3.27).

Let $\mathbf{c}_t^*$ be the optimal action at subframe t given that the joint state of subframe t is $\mathbf{s}_t^*$. The optimal scheduling policy for DP is denoted as $\mathbf{\Theta}^* = \{\mathbf{c}_1^*, \mathbf{c}_2^*, ..., \mathbf{c}_\tau^*\}$, and it is defined as a set of joint actions that should be taken across all the stages such that the total cost function is the minimized.

The DP equation presented in (3.28) is a recursive loop. At each subframe, the optimal action $\mathbf{c}_t$ is only a function of the joint state $\mathbf{s}_t$. And thus, the optimal action at

the current time does not depend on the past actions taken. Furthermore, the DP equation is backward in time, meaning that the sequence of optimal actions is determined starting from the final stage and ending at the initial stage. The solution of the DP equation works in reverse as follows. The value of $F(\mathbf{s}, \tau + 1)$ is known and equal to zero. Using (3.28), $F(\mathbf{s}, \tau)$ can be calculated as

$$F(\mathbf{s}, \tau) = \min_{\mathbf{c}} \{P(\mathbf{c}, \mathbf{s}, \tau) + 0\}. \tag{3.29}$$

As a result, $F(\mathbf{s}, \tau - 1)$ can be calculated iteratively until $F(\mathbf{s}, 1)$ is reached, which is the minimum total power transmitted during $[1, \tau]$. Thus, the optimization in (3.26) can be presented as

$$\min_{\mathbf{c}} \frac{1}{\tau} \sum_{t=1}^{\tau} P(\mathbf{c}, \mathbf{s}, t) = \min_{\mathbf{s} \in \mathbf{s}[1]} \frac{F(\mathbf{s}, 1)}{\tau}. \tag{3.30}$$

3.5 Sub-Optimal Power-Efficient Schedulers

Although DP provides global optimal solutions, it is non-causal and computationally expensive.

The BSR mechanism is responsible for informing the eNB about the current size of users' queues. The BSR is trigged when new data arrives in an empty buffer, a new data arrives that has higher priority than the one in the buffer, or when the timer for BSR (periodicBSR-Timer) expires [9]. The LTE standard defines the minimum value of the timer for BSR reporting to be five subframes [59], which is assumed in this chapter.

The DP is computationally complex for two reasons. First, it has a large number of states and actions, and second, it solves a BIP problem for each joint action. As a result, low-complexity solutions are needed for practical systems. To minimize the total transmit power and satisfy the delay constraints, three main elements should be considered while scheduling the system resources. First, data transmission at high data rates is convenient for satisfying the delay constraints. However, high data rate transmission is less power-efficient than transmitting at low data rates [8, 25]. Second, the total transmit power can be reduced by efficiently allocating the radio resources between users. As seen from (3.10), the transmit power decreases as the channel conditions improve. Third, the scheduler should

be agile to accommodate for the channel variations. For example, the scheduler may allow users to transmit high data rates when the channel is good, while transmitting low data rates in poor channel conditions. Based on the aforementioned three observations, two heuristic algorithms that solve the scheduling problem for the LTE uplink are proposed. The proposed algorithms depend on the channel quality and users' buffers lengths.

3.5.1 Maximum Transmit Power Controlling (MTPC) Scheduler

MTPC scheduler minimizes the total transmit power by preventing users from transmitting at the maximum power levels, unless it is necessary to meet the delay constraints. The objective is to maximize the total transmit DUs for all users subject to maximum allowable transmit power (MATP) levels $p_k^{max} < P^{max}$. The principal idea of the MTPC scheduler is to adapt the users' MATP levels based on their data queue lengths. The algorithm increases MATP levels for users who experience increase in their data queue length. On the other hand, MATP levels can be decreased for users who experience a decreasing queue length. In other words, the scheduler allows users who are demanding high QoS to increase their MATP, and accordingly transmit using higher rates to meet their QoS requirement. In contrast, to save power, MATPs are decreased for users who have low traffic load.

The pseudo-code in Table 3.2 describes the MTPC scheduler. The maximum number of DUs that can be transmitted over RB chunk $\mathcal{RB}_k$ while $p_k \leq p_k^{max}$ is denoted by $\Omega(\mathcal{RB}_k, p_k^{max})$. For each iteration, a single RB is allocated to the winning user who maximizes $Gain(k^*)$, where $Gain(k^*)$ is the gain user k^* achieves after granting the RB. The function $\Lambda(p_k^{max}, q_k[t])$ dynamically updates the MATP level for the next subframes, which is discuss in Section 3.5.4. The MTPC pseudo-code can be illustrated as follows

- Lines 4-9 find the best feasible RB for each user considering the contiguous allocation constraint. If a user has no allocation, the scheduler chooses the feasible RB with the highest channel gain (line 5). In case a user has been assigned one or more RBs, the scheduler chooses the feasible RB with the highest channel gain that is next to the allocated RB chunk to that user (lines 7-9). The two operations $\max|\mathcal{RB}_k|$ and $\min|\mathcal{RB}_k|$ find the highest RB number as well as the lowest RB number that

Table 3.2: MTPC scheduler.

1: $\mathcal{RB} = \{1, 2, \ldots, M\},\ \mathcal{RB}_k = \emptyset,\ \forall k \in \mathcal{K}$
2: **while** $|\mathcal{RB} \neq \emptyset|$ **do**
3: **for** $k \in K$ **do**
4: **if** $\mathcal{RB}_k = \emptyset$ **then**
5: $\mathcal{V}_k = arg \max_{m \in \mathcal{RB}} \{\alpha_{k,m}\}$
6: **else**
7: $m_U^* = \{\max|\mathcal{RB}_k| + 1\} \cap \mathcal{RB}$
8: $m_L^* = \{\min|\mathcal{RB}_k| - 1\} \cap \mathcal{RB}$
9: $\mathcal{V}_k = \max\{m_L^*, m_U^*\}$
10: **end if**
11: **if** $\mathcal{V}_k = \emptyset$ **then**
12: $Gain(k) = -1$
13: **else**
14: $Gain(k) = \Omega(\mathcal{RB}_k \cup \mathcal{V}_k, p_k^{max}) - \Omega(\mathcal{RB}_k, p_k^{max})$
15: **end if**
16: **end for**
17: **if** $\max_k Gain(k) < 0$ **then**
18: **break**
19: **else**
20: $k^* = \max_k \{Gain(k)\}$
21: $\mathcal{RB}_k^* = \mathcal{RB}_k \cup \mathcal{V}_k^*,\ \ \mathcal{RB} = \mathcal{RB} \setminus \mathcal{V}_k^*$
22: **end if**
23: **end while**
24: $q_k[t+1] = a_k[t] - \Omega(\mathcal{RB}_k, p_k^{max}[t]),\ \ \forall k$
25: $p_k^{max}[t+1] = \Lambda(p_k^{max}[t], q_k[t]),\ \ \forall k$

are assigned to user k. Consequently, $\max|\mathcal{RB}_k| + 1$ and $\min|\mathcal{RB}_k| - 1$ denote the number of the positional RBs that are adjacent to the RB chunk $\mathcal{RB}_k$.

- Lines 11-15 compute $Gain(k)$, which denotes the potential increase in number of transmit DUs by adding the best feasible RB found in lines 4-9. If the potential new allocation adds no positive gain to the user, the $Gain(k)$ is set to -1.
- Lines 17-22 determine the winning user (k^*) who achieved the maximum $Gain(k^*)$, then assign the associated $\mathcal{V}_{k^*}$ to the winning user, and take $\mathcal{V}_{k*}$ from the unallocated RBs set $\mathcal{RB}$.
- Lines 24-25 update the MATP level for all users based on their queue lengths.

Table 3.3: BWC scheduler.

$\mathcal{RB} = \{1, 2, \ldots, M\},\ \mathcal{RB}_k = \emptyset,\ \forall k \in K$
while $|\mathcal{RB} \neq \emptyset|$ **do**
  for $k \in K$ **do**
    if $\mathcal{RB}_k = \emptyset$ **then**
      $\mathcal{V}_k = arg \max_{m \in \mathcal{RB}} \{\alpha_{k,m}\}$
    else
      $m_U^* = \{\max|\mathcal{RB}_k| + 1\} \cap \mathcal{RB}$
      $m_L^* = \{\min|\mathcal{RB}_k| - 1\} \cap \mathcal{RB}$
      $\mathcal{V}_k = \max\{m_L^*, m_U^*\}$
    end if
    if $\mathcal{V}_k = \emptyset$ **then**
      $Gain(k) = -1$
    else
      $Gain(k) = \Gamma(\mathcal{RB}_k \cup \mathcal{V}_k, W_k^{min}) - \Gamma(\mathcal{RB}_k, W_k^{min})$
    end if
  end for
  if $\max_k Gain(k) < 0$ **then**
    break
  else
    $k^* = \max_k \{Gain(k)\}$
    $\mathcal{RB}_k^* = \mathcal{RB}_k \cup \mathcal{V}_k^*,\ \ \mathcal{RB} = \mathcal{RB} \setminus \mathcal{V}_k^*$
  end if
end while
$q_k[t+1] = a_k[t] - \Gamma(\mathcal{RB}_k, W_k^{min}[t]),\ \ \forall k$
$W_k^{min}[t+1] = \Xi(W_k^{min}[t], q_k[t]),\ \ \forall k$

3.5.2 Bit per Watt Controlling (BWC) Scheduler

The BWC scheduler controls the minimum acceptable Bit per Watt ratio (BPWR) for each user. Consequently, the users do not waste their power on transmission when the channel conditions are poor, and certain level of transmit power efficiency is maintained. However, when the users' queue lengths increase, the scheduler decreases the BPWR assigned to them. As a result, the scheduler pushes the users to transmit more DUs and satisfy the delay requirement at high power consumption expense.

The BPWR at subframe t is denoted by $W_k[t]$, which is defined as the number of bits transmitted per watt,

$$W_k[t] = \frac{T_k[t]}{P_k[t]}.$$

Denote $\Gamma(\mathcal{RB}_k, W_k^{min}[t])$ as the maximum number of DUs that can be transmitted over the RB chuck $\mathcal{RB}_k$ such as $W_k[t] \geq W_k^{min}[t]$, where $W_k^{min}[t]$ is the minimum acceptable BPWR at subframe t. The function $\Xi(W_k^{min}, q_k[t])$ dynamically updates the minimum acceptable BPWR, which is discussed in Section 3.5.4.

The minimum acceptable BPWR is updated based on the users queue lengths. For users who experience reduction in their queue length, the minimum acceptable BPWR is increased, otherwise, it will be decreased.

The BWC scheduler allocates RBs to users iteratively similar to the MTPC scheduler. At each iteration, one RB is allocated to the winning user k^* who maximizes $Gain(k^*)$, where $Gain(k^*)$ is the difference between the number of DUs that can be transmitted after and before assigning the potential RB to user k^* and satisfies $W_{k^*}[t] \geq W_{k^*}^{min}[t]$. The pseudo-code in Table 3.3 describes the BWC scheduler. In lines 4-9, the scheduler finds the best RB for every user, which has the highest CNR and maintains the contiguous allocation constraint. The potential gains $Gain(k), \forall k$ that are resulted from assigning the potential best RBs that are found in lines 4-9 are computed in lines 11-15. If user k has a negative gain, $Gain(k)$ is set to -1. Lines 17-19 stop the algorithm if $Gain(k) < 0, \forall k$. The winning user (k^*) is defined as the user who achieves the maximum $Gain(k^*)$ and is determined by line 20, while in line 21 the RB is assigned and the unallocated RBs set is updated. Based on the queue length, BPWRs are updated for all users in lines 24-25.

3.5.3 Complexity of the Heuristic Algorithms

The first iteration of both algorithms should find the best RB for each user. Finding the best RB, that has the maximum SNR, for a user requires M operations. Therefore, the first iteration requires $K \times M$ operations. Finding the maximum number of DUs that can be transmitted over the best RB of user k requires at most c_k^{max} operations. As constants and low-order terms don't determine the complexity order, the first iteration complexity is $O(K \times M)$. The first iteration assigns an RB to a user. In the second iteration, the number of the remaining RBs is $M - 1$. The best-case scenario is when the assigned RB in the first iteration is not a best RB for any other users. In other words, each user has a distinct best RB. In this case, no more operations are required for the next iteration because users know their best RBs, which are the same as in the previous iteration. Therefore, the best-case scenario complexity is $O(K \times M)$. Nevertheless, the worst-case scenario is when the assigned RB in the first iteration is the best RB for all users. This requires finding the second best RB for every user, which requires $O((K-1) \times (M-1))$ operations. For all M iterations, the complexity of worst-case scenario is $O(K \times M^2)$.

The heuristics algorithms can be implemented more efficiently using sorting algorithms, such as Merge sort. The SNR of the RBs are sorted for each user. Sorting M RBs requires $M \times log(M)$ operations. Consequently, $K \times Mlog(M)$ operations are required in order to sort the RBs for K UEs. Finding the best RB of a sorted array is $O(1)$. The second iteration requires only K operations to delete the assigned RB form the sorted SNR values array of every user. Thus, the complexity of iterations $2, 3, ..., M$ is $O(K \times M)$. Therefore, the algorithms complexity is determined by the first iteration which has a complexity of $K \times Mlog(M)$. As the maximum number of RBs in LTE systems is 100 [9], the worst-case complexity is $O(100K \times log(100)) \approx O(K)$ if $K \gg M$. For a scenario of low number of active users ($M \gg K$), the worst-case complexity is $O(M \times log(M))$.

3.5.4 The Controllers

For the two proposed algorithms, controllers are needed to control the MATP and BPWR in accordance with buffers' queue lengths. The MATP and BPWR are updated every l LTE subframes. Let ϑ_H and ϑ_L be positive real numbers between $[0,\ 1]$ such that $\vartheta_H > \vartheta_L$.

When the queue length of a user's buffer is greater than $\vartheta_H q_k^{max}$, the user experiences high queue length. However, when the queue length of the user's buffer is less than $\vartheta_L q_k^{max}$, the user experiences low queue length.

Tables 3.4 presents a low-complexity controller for the MATP scheduler. The MATP is set to its maximum P^{max} when the queue length of the user's buffer is greater than $\vartheta_H q_k^{max}$. Setting MATP to the maximum means that the data is transmitted at the highest possible data rates regardless of the power consumption. Therefore, the controller checks if a user experiences a critical queue length, the user is allowed to transmit using the maximum possible transmit power. If the queue length of the user's buffer is less than $\vartheta_L q_k^{max}$, the MATP decreases by δ_R. However, if the user queue length is between $\vartheta_H q_k^{max}$ and $\vartheta_L q_k^{max}$, MATP is controlled using $sign(\Delta_k)$, which is the sign of difference between $\bar{q}_{k,t-l:t-l}$ and $\bar{q}_{k,t-l:t}$, where $\bar{q}_{k,u:v}$ is the average queue length over the TTI $u : v$. If $sign(\Delta_k)$ is positive, the queue length is increased and the MATP should increase to allow the user to transmit more data and vice versa.

Tables 3.5 presents a low-complexity controller for the BPWR scheduler. The BPWR controller is similar to the MATP controller. When the queue length of the user buffers is greater than $\vartheta_H q_k^{max}$, the BPWR is set to zero. If the queue length of the user's buffer is less than $\vartheta_L q_k^{max}$, the BPWR decreases by δ_2. However, if the user queue length is between $\vartheta_H q_k^{max}$ and $\vartheta_L q_k^{max}$, the BPWR is controlled based on $sign(\Delta_k)$. The positive $sign(\Delta_k)$ values implies that the queue length increased and the BPWR level should decrease to allow the users to transmit more DUs at the expense of low bit per watt metric values. On the other hand, when the queue length decreases, $sign(\Delta_k)$ value is negative, and BPWR level decreases to force transmission at high bit per watt metric values.

3.6 Simulation Results

The performance of the proposed schedulers is evaluated using MATLAB system-level simulation based on the uplink LTE model. The DP results are solved using the solver reported in [60].

The simulated model consists of two users with different QoS requirements and arrival rates. It is assumed that both users experience identical channel conditions in terms of

Table 3.4: MATP controller.

if $q_k[l] \geq \vartheta_H q_k^{max}$ **then**
  $p_k^{max}[l] = P^{max}$
else if $q_k[t] < \vartheta_L q_k^{max}$ **then**
  $p_k^{max}[l] = \{p_k^{max}[l-1] - \delta_1, P^{max}\}$,
else
  $\Delta_k = \bar{q}_{k,t-l':t-l} - \bar{q}_{k,t-l:t}$
  $p_k^{max}[l] = \min\ \{p_k^{max}[l-1] + sign(\Delta_k) \times \delta_1, P^{max}\}$,
end if

Table 3.5: BPWR controller.

if $q_k[t] \geq \vartheta_H q_k^{max}$ **then**
  $W_k^{min}[t] = 0$
else if $q_k[t] < \vartheta_L q_k^{max}$ **then**
  $W_k^{min}[t] = 0$
else
  $\Delta_k = \bar{q}_{k,t-l':t-l} - \bar{q}_{k,t-l:t}$
  $W_k^{min}[t] = W_k^{min}[t-1] - sign(\Delta_k) \times \delta_2$
end if

the received average CNR.

Table 3.6 summarizes the considered simulation parameters. The performance of the proposed DP and heuristic schedulers is compared with each other as well as with an non-adaptive scheduler and with the greedy algorithm presented in [25]. The non-adaptive scheduler is similar to the MTPC scheduler but the MATP levels are fixed to P^{max}. The non-adaptive scheduler transmits at the maximum possible data rate regardless of the channel condition or the buffer queue length. As a matter of fact, the non-adaptive scheduler can be seen as the BWCA scheduler with $W_k^{min}[t] = 0,\ \forall t,$ and $\forall k$.

Fig. 3.6 shows the average user transmit power per TTI. As the average channel gain increases, the average transmit power decreases for all the considered schedulers and users scenarios. As shown in (3.10), the transmit power is inversely proportional to the channel gain. Therefore, increasing the average channel gain reduces the transmit power. The DP consumes the least average transmit power compared to the other schedulers. The DP scheduler provides the optimal solution, in which both the channel gains and the data arrival

Table 3.6: Parameter settings of the uplink LTE model.

Parameter	Setting	Parameter	Setting
Number of RBs (M)	4	Number of users (K)	2
P^{max}	23 dBm (200mw)	λ_1	1.5
c^{max}	5	λ_2	2.5
ϑ_L	0.15	ϑ_H	0.75
δ_1	150	δ_2	0.02
q_1^{max}	60	q_2^{max}	80
l	5	τ	1000

rates are known for the whole scheduling time before the scheduling begins. Since it has higher arrival rate, UE2 transmits more DUs and consumes higher power than UE1. Since $\lambda_2 > \lambda_1$, the average transmitted data of UE2 is higher than that of UE1. Consequently, UE2 consumes more power than UE1. The performance of MTPC scheduler outperforms the performance of BWC. The non-adaptive scheduler consumes the highest power among the schedulers. On average, MTPC and BWC schedulers consume 0.6 % and 0.70 % of the power consumed by the non-adaptive scheduler, respectively.

Fig. 3.7 and Fig. 3.8 present the average queue length of UE1 and UE2, respectively. Although the DP scheduler experiences the highest average queue length, it guarantees that no delay violation occurs. On average, for the MTPC and BWC schedulers, the user queue lengths are 57% and 80% less than the queue length resulting by the DP solution, respectively. The non-adaptive scheduler experiences average queue length 7.5 times less than the DP scheduler. The non-adaptive scheduler tends to transmit data at the possible maximum rates, which reduces the average queue lengths at the expense of higher transmit power.

Fig. 3.9 and Fig. 3.10 show the PDFs of the queue length of UE1 and UE2, respectively. The PDFs of the non-adaptive and the BWC schedulers are similar and almost confined between 0 and $0.3\, q_k^{max}$. The queue length PDFs for the MTPC are similar to the PDF for the optimal solution. The PDFs for the MTPC and BWC schedulers drop when the queue lengths approach to $\vartheta_H q_k^{max}$, until they reach zero for queue length over than q_k^{max}. When a queue length reaches $\vartheta_H q_k^{max}$, both schedulers allow the user to transmit at the highest rate possible, which results in zeroing the queue length PDFs for values over

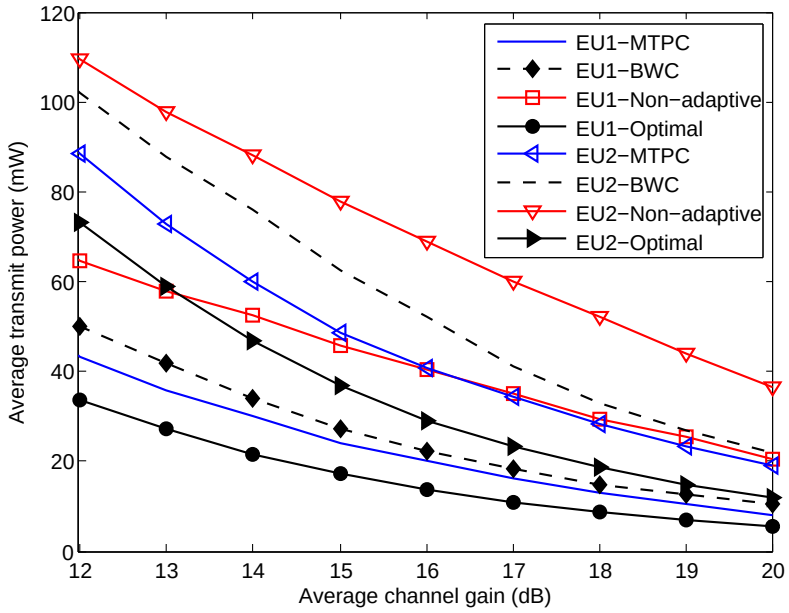

Figure 3.6: Average transmit power per TTI.

than $\vartheta_H q_k^{max}$.

However, the DP scheduler knows the arrival rate and the channel gain before scheduling, and it manages the delay to be less than the maximum allowed delay and reduces the total power transmit. Therefore, the queue length PDFs for the DP scheduler show the highest values among the schedulers between around the maximum allowed delay value. It is worth mentioning that unstressed system is considered and more investigations on the utilization of the system is kept for future work.

3.6.1 Large-Scale Scenario

In this subsection, the iterative algorithms are evaluated for a large-scale scenario. As the DP complexity increases exponentially with the problem size, the optimal solution is not included for the large-scale scenario. Table 3.7 summarizes the simulation parameters for the large-scale scenario.

Fig. 3.11 and Fig. 3.12 show the average user transmit power per TTI and the average queue length (DU), respectively. As the number of users increases, the competition for the

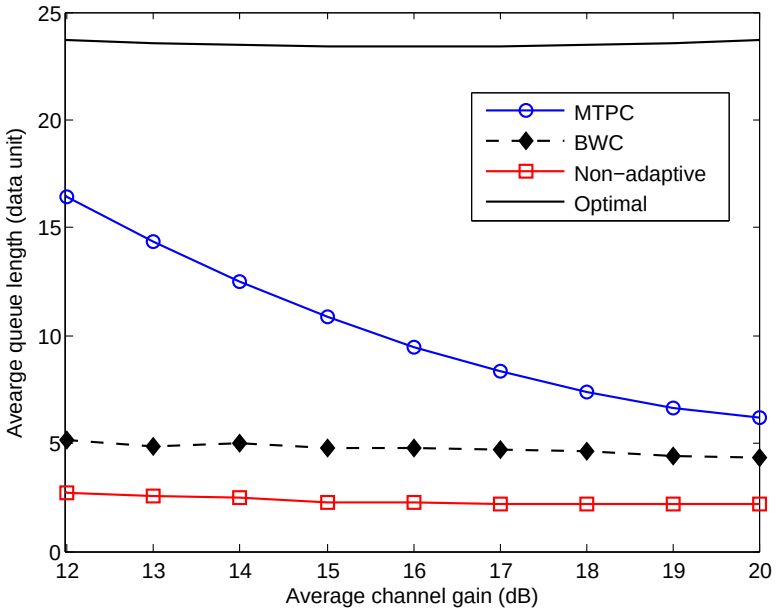

Figure 3.7: Average queue length for UE1.

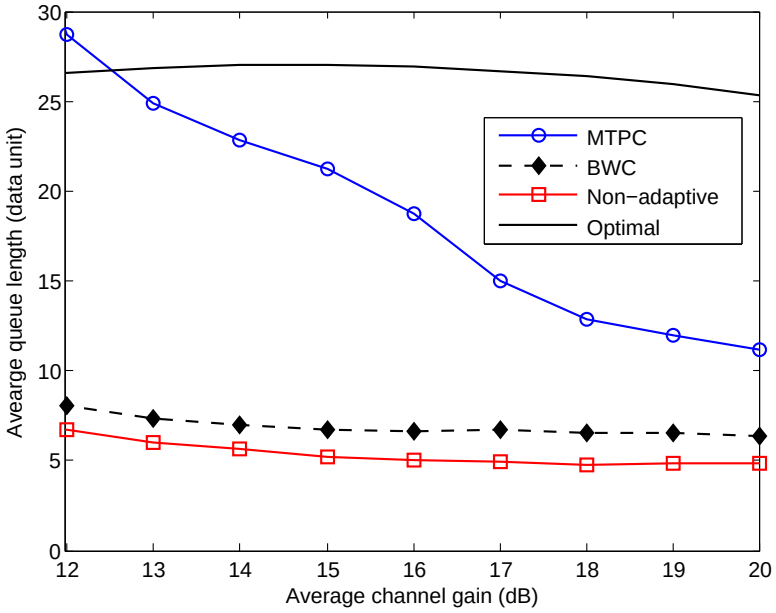

Figure 3.8: Average queue length for UE2.

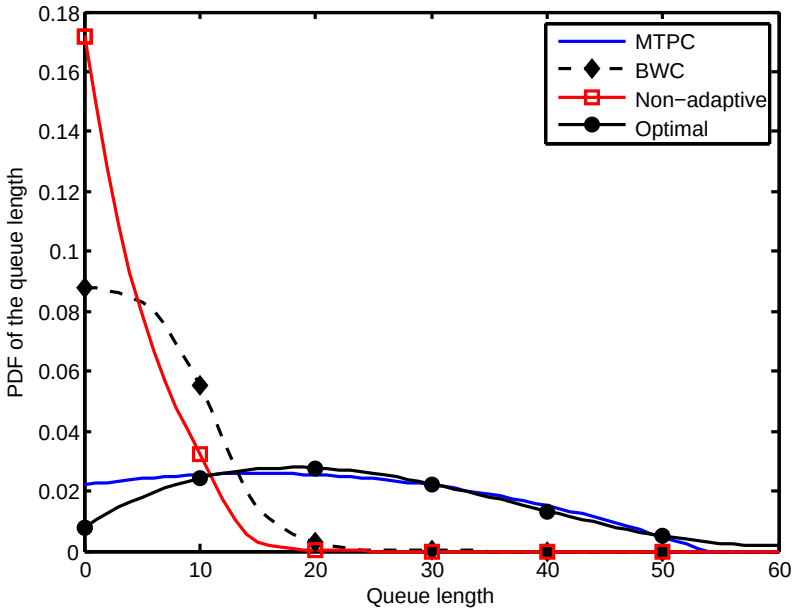

Figure 3.9: Probability density function of the queue length for UE1 (average CNR=13 dB).

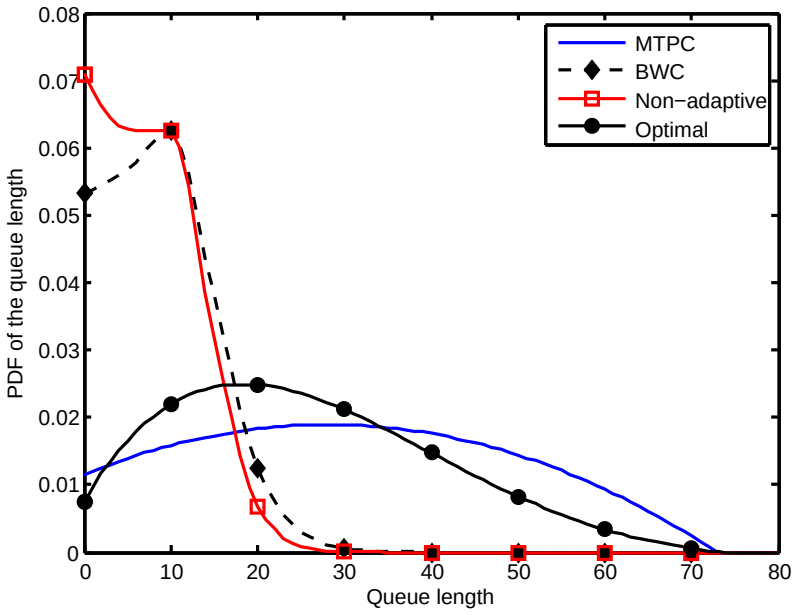

Figure 3.10: Probability density function of the queue length for UE2 (average CNR=13 dB).

Table 3.7: Parameter settings of the large-scale scenario.

Parameter	Setting	Parameter	Setting
Number of RBs (M)	100	Average CNR	15 dB
$\lambda_k, \forall k$	5	$q_k^{max}, \ \forall k$	200

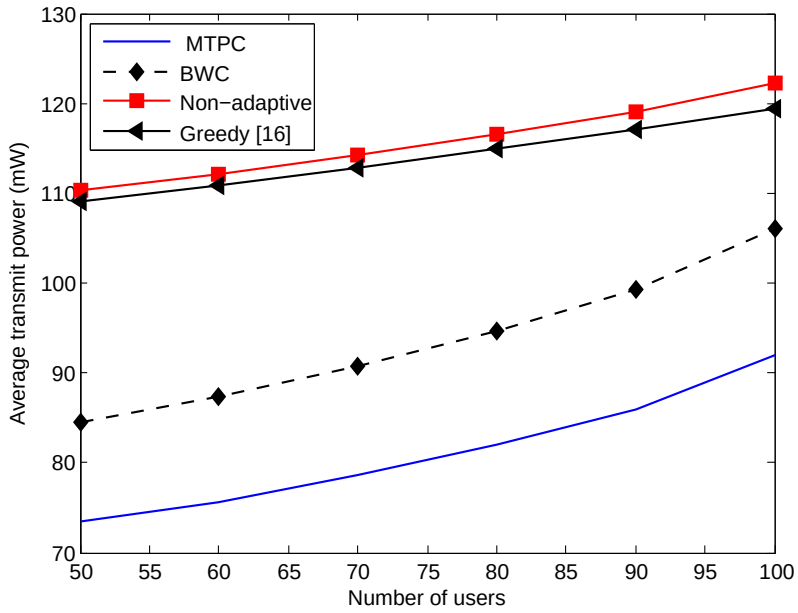

Figure 3.11: Average transmit power per TTI for the large-scale scenario.

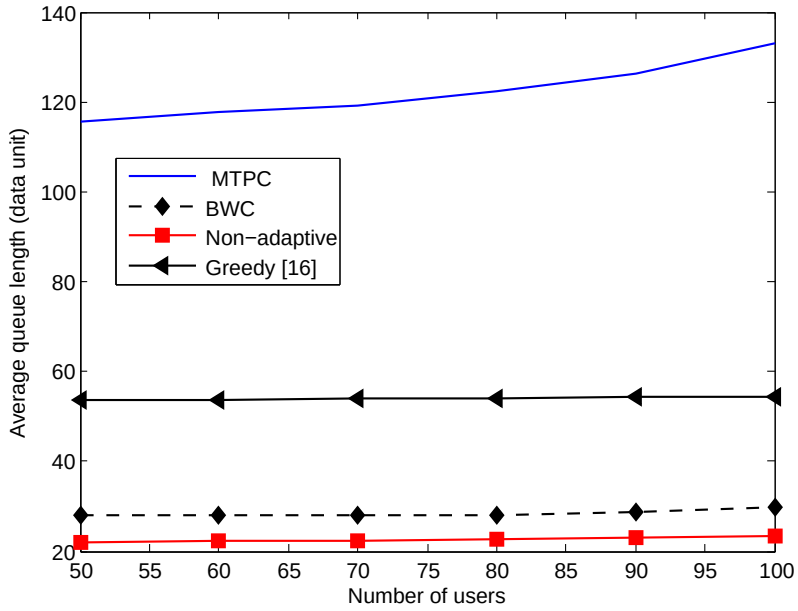

Figure 3.12: Average queue length for all users for the large-scale scenario.

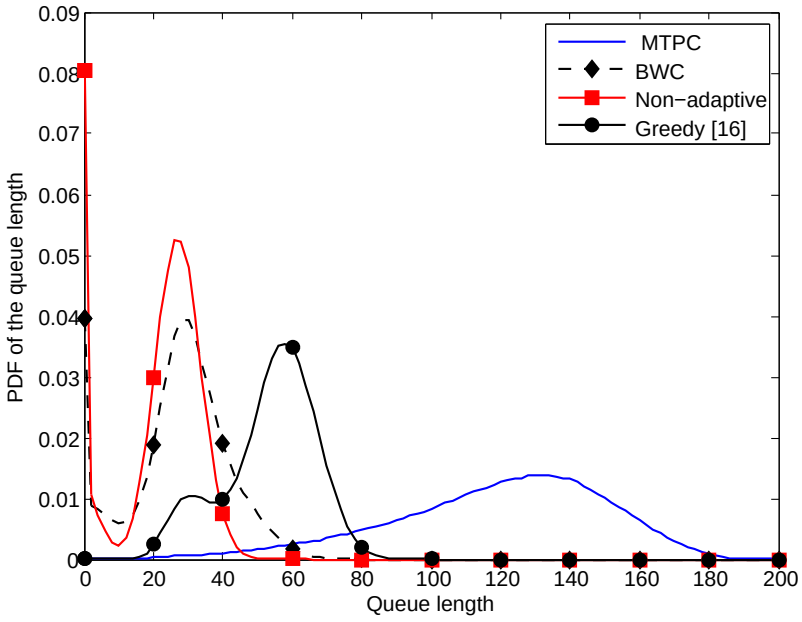

Figure 3.13: Probability density function of the queue length for number of users 70.

radio resources increases leading to transmitting more power and growth in queue lengths for all schedulers.

The average queue length is inverse proportional to the power consumed. By using MTPC scheduler, users consume the least power compared to using BWC, greedy [25], and non-adaptive schedulers. However, users experience the highest average queue lengths. However, by using the non-adaptive scheduler, users have the least queue length but consume the most power among the schedulers.

Fig. 3.13 shows the PDFs of the queue lengths averaged on all users. For all schedulers, the average queue length is lower than the maximum allowed (200 DUs), which means no delay violation has is occurred.

Fig. 3.14 and Fig. 3.15 show the average transmit power and queue length for the MTPC and BWC schedulers across different values of l. As l increases, the power consumption slightly increases and the average queue length decreases for both schedulers. Note that the MATP and BPWR are updated once every l subframes. Updating the MATP and BPWR at higher rates allows the schedulers to save more power. However, updating the MATP and BPWR at lower rates decreases the delay as a result of transmitting at higher data rate at the expense of the power consumption.

3.7 Chapter Summary

In this chapter the problem of minimizing transmit power subject to delay constraints for the LTE uplink systems is considered. The optimal solution based on a dynamic programming approach is derived. The scheduling problem is divided into stages, each stage has a number of states based on the queue length of the users' buffers. Actions are taken at each stage based on the channel quality and the queue length. Binary integer programming is invoked to find the minimum cost for each action. The minimum action costs results from optimally allocating the resource blocks to the users. After finding the minimum cost for each action, the optimal actions are derived by solving the dynamic equation backward starting from the last stage until the first stage. Consequently, two online schedulers that solve the scheduling problem with comparable power consumption are proposed. The

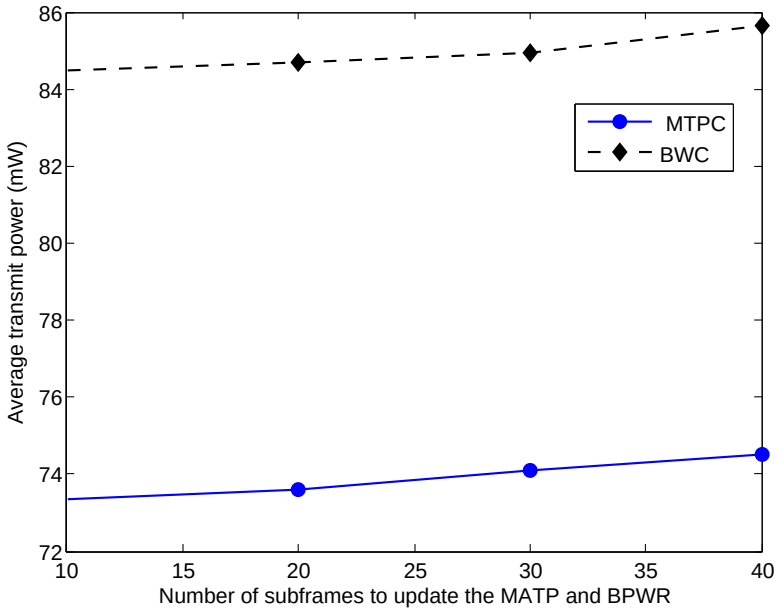

Figure 3.14: Average transmit power for all users per TTI for different l.

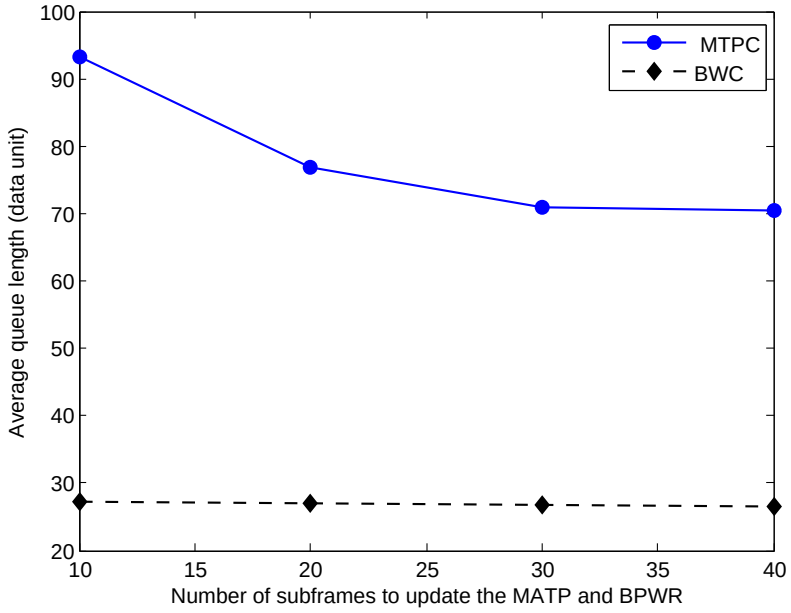

Figure 3.15: Average queue length for all users for different l.

schedulers are based on low computational complexity heuristic algorithm. The first scheduler adapts the maximum allowable transmit power for every user according to the queue length of the user's buffer. The scheduler prevents user to transmit at high power levels when the buffer queue length is low. Whereas, the second scheduler controls the minimum acceptable bit per watt ratio (BPWR) for each user. Users can only transmit if they experience BPWR greater than a minimum acceptable BPWR. For the online schedulers, controllers are derived to adapt the MATP and BPWR levels for users based on the queue length of their buffers. The online schedulers performances are compared with the optimal and existing schedulers in terms of transmit power and delay.

Chapter 4
Wireless Resource Virtualization: Opportunities, Challenges, and Solutions

4.1 Introduction

In the last decade, the demand for mobile data services has experienced a substantial growth that provoked the mobile network operators (MNOs) to search for novel solutions to satisfy the surging demand while increasing the average revenue per user. Mobile data traffic forecasts provided by companies such as Cisco, estimated 11-fold increase in the global mobile data traffic between 2013 and 2018 [4]. Thus, MNOs need to stretch the capacity of their mobile networks into a new horizon to satisfy the expected demand. As the revenues generated do not commensurate with the traffic growth, operators should increase their capacity while being extremely cautious with their investment. They must efficiently utilize the scarce and highly expensive wireless resources. For example, in 2011, the FCC held Auction 92 in order to sell spectrum licenses in the 700 MHz band. The seven winners paid $19.8 billion for 16 licenses. On the other hand, recent spectrum utilization measurements have shown that wireless resources are underutilized in many cases, which urges the call for action to create innovative solutions for the problem of underutilized wireless resources.

Wireless resource virtualization (WRV) is gaining a remarkable interest by the industry as a promising solution to address the problems of spectrum scarcity and resources utilization inefficiency. WRV enables network operators to create multiple logical networks on a single physical substrate, thus yielding better efficiency in terms of energy consumption and hardware utilization as shown in Fig. 4.1, where a single base station (BS) is shared among three MNOs. Sharing BS hardware components as well as spectrum resources are of great interest to MNOs. A recent study from ABI Research [61] shows that a worldwide

active infrastructure sharing over a period of five years can save up to $60 billion in capital expenditure (CAPEX) and operating expenditure (OPEX).

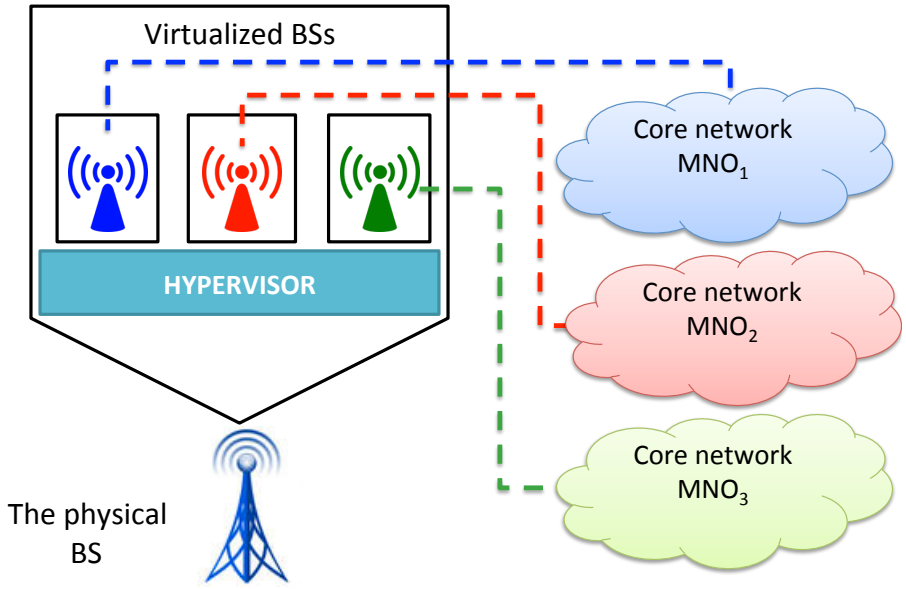

Figure 4.1: Base station virtualization.

While virtualization of wireless networks has recently received an increased attention, virtualization in wired networks began over twenty years ago. Although general concepts of network virtualization remain the same, solutions from wired networks cannot be directly applied to wireless networks. The main reason is that links in wired networks are reliable, physically isolated from each other, and have a constant bandwidth. Hence, the physical layer can often be ignored without major impact on the network performance. The same approach cannot be applied to wireless networks. Wireless links are less reliable, suffer from interference, and have a fluctuating capacity depending on the channel quality. Therefore, the physical layer of wireless networks profoundly affects the network performance and should be taken into consideration during the virtualization process. Furthermore, the resources allocation process in wireless networks faces additional challenges as a result of the dynamic nature of the wireless channel, wireless network topologies, mobility of users and fast variation of Quality of Service (QoS) requirements of on-the-go applications. These challenges mandate the presence of a hypervisor that is capable of monitoring the resources allocation to different networks, coping with the rapid varia-

tions of the wireless channel and users' demands, and assuring the adherence to the sharing agreement between the MNOs.

The remaining part of the chapter is organized as follows. Sections 4.2 and 4.3 discuss the main benefits and challenges of adopting WRV. In Section 4.4, possible solutions for WRV are investigated. Section 4.5 proposes efficient fair low-complexity solution for wireless resource virtualization in LTE networks, and Section 4.6 concludes the chapter.

4.2 Main Benefits of WRV

WRV offers pivotal benefits to the wireless market as well as the environment. The main benefits of WRV can be divided into three groups as follows.

4.2.1 Economic Sharing of Investment and Cost Reduction

The most immediate benefit of WRV is the considerable reduction of the CAPEX and OPEX. By consolidating their equipments and spectrum licenses, MNOs are able to enhance their network coverage and to alleviate the capital investment needed for launching new services, which minimizes the time-to-market of new innovations. For example, an MNO does not have to solely own a certain band through expensive and uncertain bedding systems. It needs simply to get into a sharing agreement with an existing MNO who owns the required band.

When multiple network operators jointly construct a network, hardware equipment such as antennas, cooling systems, and towers can be shared, which reduces the operational costs and power consumption. Moreover, resources sharing overcomes the capital shortage related to new infrastructure investments, resulting in shorter network deployment periods and expedited time-to-market process. Additionally, WRV enables the incorporation of spectrum pooling techniques to improve the spectrum utilization and increase the total network capacity [62]. Therefore, a great economic potential is created for reaping the benefits of the growing wireless market, and it enables business innovation in mobile services and applications.

4.2.2 Collaborative Business Models

Governments assign spectrum licenses through competitive mechanisms such as spectrum auctions or competitive tender, which is commonly referred to as beauty contests. Because the spectrum available for transmission is limited, a small number of MNOs are usually dominating the wireless market, which increases the cost of services provided to the end users. Wireless virtualization enables the creation of a virtual MNO (vMNO) that purchases the spectrum usage rights from an MNO and provides wireless services to the end customers under its own brand-name. vMNO is a new business model that enables fair competition that benefits the end customers. Furthermore, it improves resource utilization by selling underutilized network capacity.

4.2.3 Environmental Benefits

Energy consumption in wireless networks is a growing concern for MNOs due to the increasing energy prices and the environmental consequences. The information and communication technology industry currently consumes about 3% of the global energy and emits 2% of the global Carbon dioxide (CO_2) [63]. Such numbers may rise sharply if each new MNO has to install new infrastructure for each new network.

While many works focus on virtualizing the core network [64], a significant portion of the energy consumed by cellular networks is used by the radio access network (RAN) subsystem. Thus, WRV can increase the energy efficiency of cellular networks because sharing equipment and sites between multiple MNOs reduces the number of new access nodes required by new market players, or by existing operators who want to extend the coverage of their networks. Therefore, the energy consumed and the expansion of the wireless communications sector will not be growing at the same rates. Or, the current MNOs can utilize the WRV to reduce the current energy consumption and the CO2 emissions, which promotes the green radio communication. Recently, energy consumption has become a key element in designing wireless systems and building green networks. Although most previous work has focused on power consumption at the user equipment (UE) end to extend the mobile battery life, more recent efforts have been concentrated on the wireless network energy consumption due to its economic and environmental benefits [65].

4.3 Operational and Business Challenges of WRV Deployment

In spite of the potential benefits and cost-savings of WRV, two main operational and business challenges prevent widespread deployment of WRV:

4.3.1 The Risk of Market Share Loss and Anti-Competitive Practices

While WRV facilitates network deployment, accelerates the penetration of start-up operators and boosts the market competition, it exposes the well-established large operators, to the risk of market share loss. The large operators hope to achieve cost savings through infrastructure sharing with competitors. However, it raises the fear that relaxing the requirements for new players to enter the mobile market may dilute their market shares. Consequently, large network operators will attempt to outline the WRV process to empower themselves to control the market and reduce the success chances of smaller operators.

4.3.2 Independence of Services with RAN-Sharing

The operators' ability to control the network entities is essential for innovation and provision of new services. While WRV consolidates hardware and resources of multiple MNOs, it limits the operators' independence and slows down services deployment since other operators might be influenced by particular decisions. Therefore, deducing new strategies to enable the deployment of particular services for some operators in the context of WRV is still required and remains a challenging aspect for the WRV process.

4.4 Scope of Virtualization and Depth of Sharing

Wireless resource sharing can be achieved at different levels. Two main scopes can characterize the depth of sharing, namely, passive sharing and active sharing.

4.4.1 Passive Sharing

In passive sharing, operators share passive infrastructure entities such as radio masts and towers, power supply, air conditioning sites and building premises. Passive sharing has become popular since 2000 because it reduces initial investment and saves operating costs. Since no active operational coordination between operators is required, passive sharing is straightforward to apply, which makes it the most widespread sharing scope, particularly in low-density areas. For example, 40% of Orange sites in the rural areas of France are subject to passive sharing. Another example of passive sharing is the agreement between Vodafone and Telefonica in 2009 where they announced a wide-ranging agreement to share mobile networks' infrastructure in the UK, Spain, Ireland, and Germany to cope with the demand for mobile broadband data services and to save millions of Pounds in costs [66]. However, the limitation of passive sharing is that no network capacity improvement is expected.

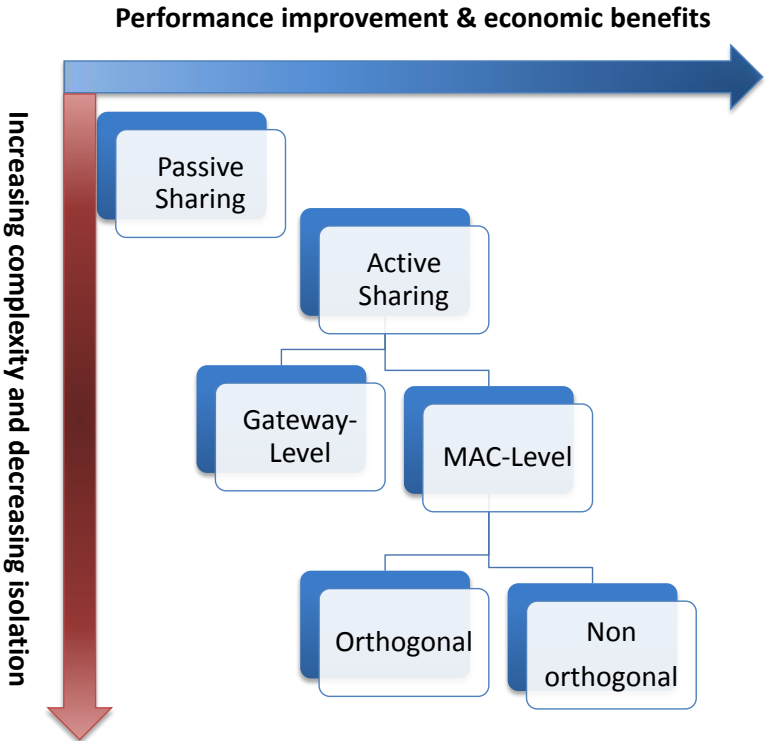

Figure 4.2: Scope of virtualization.

4.4.2 Active Sharing

In active sharing, network active elements are included such as electronic infrastructure, transceivers, antennas, spectrum, and fiber optic networks. Furthermore, the sharing can be extended to reach core network components such as mobility management entity (MME). Active infrastructure sharing provides more savings in CAPEX and OPEX as compared to passive sharing. Therefore, it is receiving a proliferating attention by wireless operators and standards development organizations. The 3GPP LTE standard supports two active sharing configurations as depicted in Fig. 4.3, multi-operator core network (MOCN) and gateway core network (GWCN). In MOCN configuration, each MNO has its own core network. However, GWCN extends the sharing to the core network, where MNOs can also share the MME of the core network, which enables additional cost savings but reduces the level of isolation and dependency between MNOs. Each core network operator is identified by a distinguished public land mobile network identification (PLMN-ID). Shared cells broadcast the PLMN-IDs of the MNOs that share the network. Each UE access the shared network by selecting one of available PLMN-IDs. Active sharing can be achieved at different levels. Deeper level of sharing results in higher CAPEX and OPEX reduction, but increases the deployment complexity and limits the freedom of operators to manage and operate the network. Active sharing solutions can be classified into two groups based on the level of sharing, which are outlined in the following two subsections; Gateway-Level solution and MAC-Level solution.

4.4.2.1 Gateway-Level Solution

The work reported in [61] proposes a gateway-level solution that enables multiple MNOs to share one single physical BS without modifying its MAC schedulers. The main advantage of such solution is the facilitation of immediate deployment of RAN sharing since no major modifications are required to existing BSs. However, the resource allocation decisions are invisible to the gateway, which degrades the performance of the sharing process. In such solutions, the virtualization happens by reshaping the traffic at the gateway level. For the downlink, the gateway controls the number of packets sent to the BS for each operator. While sending the traffic to the BS, the gateway considers sharing requirements such

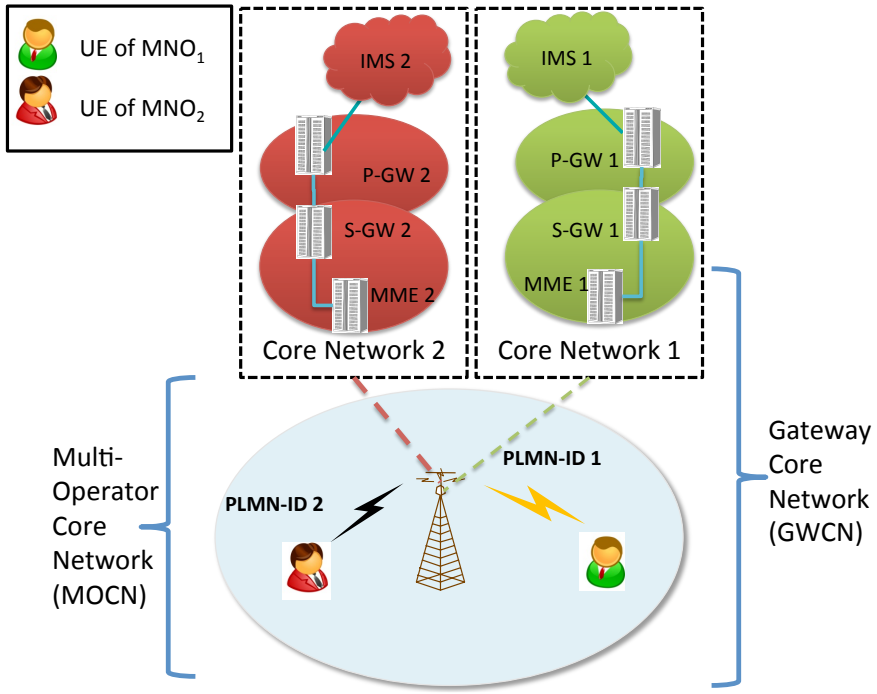

Figure 4.3: LTE sharing configuration options.

as efficient resource utilization and isolation between different operators. For the uplink, controlling the resource allocation is not straight-forward because the traffic is generated by the end users, which cannot be directly controlled by the gateway. This can be solved by creating a feedback channel between the BS and the gateway. The gateway sets an upper-bound for the average data rate for each data flow. The gateway controls the uplink sharing by tuning the flows' upper bound values.

4.4.2.2 MAC-Level Solution

virtualizing an LTE BS at the MAC level is tackled in [62] where full control of the internal scheduler of the BS is accessible. The main motivation is to provide the most efficient sharing of the wireless resources between users who belong to different MNOs. The scheduler considers various perspectives while assigning the resources to users such as the quality variation of the wireless channel, satisfying the MNOs' service-level agreements (SLAs), efficient utilization of the wireless resources, providing tight isolation between MNOs who are sharing the same physical substrate, and enabling MNOs to implement

their own custom scheduling policies that fit their service requirements and business models. This solution gives the maximum efficiency of resource utilization as the resource allocation decisions are fully controlled at the BS level. Another LTE air interface virtualization scheme is proposed in [67], where a hypervisor is added on top of the physical resources. This hypervisor is responsible of virtualizing the evolved node-B (eNB) into a number of virtual eNBs that are then used by different MNOs. It was shown that the capacity gains can be achieved by sharing the spectrum resources between different MNOs. More practical scenarios are studied in [68] where MNOs share multiple eNBs, with the sharing process managed by the hypervisor. Enhancements such as load balancing and safety margins were investigated. A flow-level virtualization scheme of wireless resources on BSs in WiMAX network was proposed and evaluated in [69]. The BS radio resources are sliced between different flow groups, and the solution of [69] enables customized flow scheduling per slice, where each slice can be treated as a virtual MNO that supports a set of flows.

Resources in MAC-Level solution can be even more efficiently utilized if we use Non-Orthogonal sharing where a certain frequency can be used more than once in the same cell.

Orthogonal Sharing: multiple operators share their spectrum without introducing mutual interference. Therefore, a radio block is assigned only to one operator in any given time slot. Sharing can be (a) full: where the MNOs aggregate and share their entire spectra, or (b) partial: where MNOs prefer to have sole ownership on a certain band in order to satisfy QoS guarantees for their customers. The authors of [70] show that the overall theoretical throughput gain of orthogonal sharing is about 12%. Although it is not a substantial gain, it comes almost at no cost. It is just a matter of taking advantage of the asymmetry of the loads between the MNOs. Orthogonal sharing might require BS with extra capabilities to be able to accommodate a larger spectrum and a high data backbone connection in order to convey the increased traffic.

Non-orthogonal Sharing: in this scenario, a frequency band can be allocated simultaneously to two operators. This model offers much higher spectrum efficiency gains, yet it increases the complexity of the system due to the fact that interference reduction or cancellation techniques should be incorporated. The enablers to manage the interference between

different BSs to improve the overall system performance are discussed in [70]. The main enabler is called transmit beamforming where a BS with multiple antennas can steer the transmission power towards a certain destination using the appropriate scaling of the transmitted signal in each antenna. Hence, interference is managed by spatial separation rather than time or frequency separation as in TDMA or FDMA. In order to take advantage of beamforming, the MNOs have to exchange channel state information via an appropriate feedback interface.

4.5 Efficient Fair Low-Complexity Scheduler for WRV

A low-complexity scheduler that virtualizes the wireless resource blocks (RBs) at a single eNB and shares them between users of different MNOs is presented in this section. It is worth mentioning that network-wide virtualization will be discussed in Chapter 5. The scheduler's target is to maximize the throughput of all users while maintaining average RB access probabilities (RAPs) between different MNOs and access proportional fairness (APF) between users of each MNO. To facilitate readability, Table 4.1 summarizes the notations frequently used throughout the chapter.

Table 4.1: Summary of the most significant notation.

Symbol	Meaning
K, k	Total number of UEs, UE index
B, b	Total number of RBs, RB index
M, m	Total number of MNOs, MNO index
K_m	Number of users who subscribe to MNO m
$\mathcal{K}_m$	Set of users who belong to MNO m
r_m	RAP of MNO m for one RB
γ_k	Instantaneous received SNR of user k
$\mathcal{B}_k$	Set of RBs allocated to user k
$\bar{A}_1$	Average number of RBs accessed by MNO m
$\eta_{k,b}$	Normalized instantaneous SNR on RB b seen by user k
β	Binary number indicator
$S_{k,b}$	Transport block size of user k over RB b

4.5.1 System and Channel models

Consider an LTE downlink system where B RBs are shared between M MNOs each of which serves $K_m, m = \{1, 2, ..M\}$ active users who have backlogged data at an eNB. A sharing agreement specifies the RB access probabilities (RAPs) for MNOs. The RAPs are assumed to be identical for all RBs. Define r_m as the RAP of MNO m for one RB, the RAPs for MNO m can be expressed as

$$A_m = B \times r_m. \tag{4.1}$$

The finest resource granularity for transmission in LTE is one subframe in time and one RB in frequency. The LTE subframe has a duration of 1 ms. Each RB consists of 12 successive subcarriers and has a bandwidth of 180 kHz. The channel is modeled as a quasi-static frequency-flat Rayleigh fading channel. The channel gain is constant over one RB bandwidth and one subframe time interval, but changes independently over successive RBs and subframe time intervals. Furthermore, users are assumed to experience independent fading. Equal power allocation is assumed where the total transmit power at the eNB is equally divided between all RBs [71]. Furthermore, perfect channel state information is assumed to be available at the eNB for all users. At the receiver side of user k, the instantaneous received SNR γ_k is assumed to be exponentially distributed with average $\bar{\gamma}_k$.

As assumed in Section 2.3, three symbols in each frame are assigned to uplink physical control signalling. Therefore the transport block size at time slot t of user k over RB b can be expressed as

$$S_{k,b}[t] = \lfloor 132\, \Phi(\gamma_b[t]) \rfloor \tag{4.2}$$

where $\Phi(\gamma_b[t])$ is the modulation and coding schemes (MCS) efficiency used over RB b, and $\lfloor . \rfloor$ denotes the floor function. The selection of MCS is performed using tables that map the received SNR to a MCS as described in Section 2.3.

By denoting the set of RBs allocated to user k by $\mathcal{B}_k[t]$, the total transport block size

of user k can be calculated as

$$S_k[t] = \sum_{b \in \mathcal{B}_k[t]} \lfloor 132\ \Phi(\gamma_b[t]) \rfloor\ . \tag{4.3}$$

4.5.2 Problem Formulation

The scheduler's objective is to maximize the transmission rate of all users subject to long-term APF between users and MNOs. The scheduling problem can be formulated as

$$\max \sum_{t=l}^{l+T} \sum_{m=1}^{M} \sum_{b=1}^{B} \sum_{k \in \mathcal{K}_m} S_{k,b}[t]\ \beta_{k,b}[t] \tag{4.4a}$$

subject to

$$\sum_{k=1}^{K} \beta_{k,b}[t] \leq 1,\ \forall\, b = 1, 2, ..., B \tag{4.4b}$$

$$\bar{A_1} = B\ A_1,\ \bar{A_2} = B\ A_2,\ \cdots,\ \bar{A}_M = B\ A_M \tag{4.4c}$$

$$a_u = a_v = \bar{A}_m / K_m, \forall\, (u, v) \in \mathcal{K}_m, \forall\, m = 1, 2, ..., M \tag{4.4d}$$

where $\mathcal{K}_m$ is the set of users who belong to MNO m, K is the total number of users, $\bar{A_1} = E_t \left[\sum_{k \in K_m} \sum_{b=1}^{B} \beta_{k,b}[t] \right]$ is the average number of RBs accessed by MNO m, $E_t[.]$ is the expectation with respect to t, and $\beta_{k,b}[t]$, is a binary number indicator which is set to one if user k is granted RB b at time slot t, and zero otherwise. Constraint (4.4b) prevents allocating RBs to more than one user. Constraint (4.4c) maintains APF between MNOs while constraint (4.4d) maintains APF between users by forcing them to have equal long-term average RAP.

Finding the global optimal solution for (4.4) is computationally expensive due to the large size of the search space. In addition, it may require the knowledge of future channel gains for all users. To reduce the complexity, the weighted normalized SNR based scheduler is proposed as a practical solution that offers guaranteed performance gain and convergence. Consequently, the optimal weights that maximize the cell throughput and maintain APF are computed. APF between users of each MNO can be attained using

normalized SNR selection criterion [72]

$$\max \sum_{k \in \mathcal{K}_m} \sum_{b=1}^{B} \eta_{k,b}\, \beta_{k,b} \tag{4.5a}$$

subject to

$$\sum_{k=1}^{K} \beta_{k,b} \leq 1, \ \forall\, b = 1, 2, ..., B \tag{4.5b}$$

where $\eta_{k,b} = \gamma_{k,b}/\bar{\gamma}_k$ is the normalized instantaneous SNR on RB b seen by user k. However, normalizing the SNR for all users imposes APF between all MNOs users, but it does not necessarily satisfy (4.4c). To maintain (4.4c), the normalized SNR of each MNOs users should be weighted differently. Thus, the weighted normalized SNR scheduler can be formulated as

$$\max \sum_{m=1}^{M} \sum_{b=1}^{B} \sum_{k \in \mathcal{K}_m} \alpha_m \eta_{k,b}\, \beta_{k,b} \tag{4.6a}$$

subject to

$$\sum_{k=1}^{K} \beta_{k,b} \leq 1, \ \forall\, b = 1, 2, ..., B \tag{4.6b}$$

where α_m is the weighting factor of MNO m, and $\sum_{m=1}^{M} \alpha_m = 1$. Therefore, user $k \in \mathcal{K}_m$ is assigned RB b if the following equation holds

$$k = \arg \max_{k \in \mathcal{K}_m} \alpha_m \eta_{k,b} \tag{4.7}$$

where $\alpha_m \eta_{k,} \triangleq \zeta_{k,b}$ is the weighted normalized SNR of RB b seen by user $k \in \mathcal{K}_m$. For independent and identically distributed Rayleigh-faded users, the PDF of $\zeta_{k,b}$ given that user k is assigned RB b ($\zeta_{k,b} > \zeta_{i,b}, \forall i \in \mathcal{K}, i \neq k$) is given by [72]

$$f_{\zeta_{k,b}}(x) = \frac{1}{\bar{\zeta}_{k,b}} \mathrm{e}^{-x/\bar{\zeta}_{k,b}} \prod_{i=1, i \neq k}^{K} 1 - \mathrm{e}^{-x/\bar{\zeta}_{i,b}}, \tag{4.8}$$

where $\bar{\zeta}_{k,b} = E[\zeta_{k,b}]$. By assuming that $\bar{\zeta}_{k,b}$ is equal for all users who belong to the same

MNO, then (4.8) can be written as

$$f_{\zeta_{k,b}}(x) = \frac{1}{\bar{\zeta}_{k,b}} \mathrm{e}^{-x/\bar{\zeta}_{k,b}} \Theta_{k,b}^{K_m-1}(x) \prod_{i=1, i\neq m}^{M} \Theta_{k,b}^{K_i}(x) \tag{4.9}$$

where $\Theta_{k,b}^{K_m}(x) = (1 - \mathrm{e}^{-x/\bar{\zeta}_{k,b}})^{K_m}$. Thus, user k average RAP to a single RB is given by

$$a_{k,b} = \int_0^{\infty} f_{\zeta_{k,b}}(x) dx \tag{4.10}$$

using the binomial expansion, $\Theta_{k,b}^{K_m}(x)$ can be expressed as

$$\Theta_{k,b}^{K_m}(x) = \sum_{n_m=0}^{K_m} \binom{K_m}{n_m} (-1)^{n_m} \left(\mathrm{e}^{-n_m x/\bar{\zeta}_{k,b}}\right). \tag{4.11}$$

By substituting (4.11) and (4.9) in (4.10), $a_{k,b}$ can be written as

$$a_{k,b} = \frac{1}{\bar{\zeta}_{k,b}} \sum_{n_1=0}^{K_1} \sum_{n_2=0}^{K_2} \cdots \sum_{n_m=0}^{K_m-1} \cdots \sum_{n_M=0}^{K_M} \Psi_{\mathbf{n}} \int_0^{\infty} \mathrm{e}^{-x\Phi_{\mathbf{n}}} dx \tag{4.12}$$

where $\mathbf{n} = [n_1, \cdots, n_M]$,

$$\Phi_{\mathbf{n}} = \frac{1}{\bar{\zeta}_{k,b}} \left[1 + \sum_{i=1}^{M} n_i\right] \tag{4.13}$$

and

$$\Psi_{\mathbf{n}} = \binom{K_m - 1}{n_m} (-1)^{K-1} \prod_{i=1, i\neq m}^{M} \binom{K_i}{n_i}. \tag{4.14}$$

Finally, $a_{k,b}$ can be written as

$$a_{k,b} = \frac{1}{\bar{\zeta}_{k,b}} \sum_{n_1=0}^{K_1} \sum_{n_2=0}^{K_2} \cdots \sum_{n_M=0}^{K_M} \frac{\Psi_{\mathbf{n}}}{\Phi_{\mathbf{n}}}. \tag{4.15}$$

Consequently, user k of MNO m is assigned $a_k = a_{k,b} \times B$ RBs on average. Therefore,

the RAP for MNO m is

$$A_m = K_m \times a_k. \tag{4.16}$$

It is worth noting that for fixed number of users, the RAPs for the MNOs only depend on $\bar{\zeta}_{k,b} = E[\zeta_{k,b}] = \alpha_m, \forall k \in \mathcal{K}_m$. Therefore, finding the proper weights can maintain (4.4c).

Although it is assumed that all users subscribed to a particular MNO are equally weighted to maintain (4.4d), users of a single MNO may have different priorities. Therefore, they should have different $\bar{\zeta}_{k,b}$ to access different average number of RBs. In such scenarios, the MNO, denoted as parent-MNO (PMNO), is segmented into a number of groups denoted as child-MNO (CMNO). Each CMNO segmentation is assigned different weight based on the users' priorities. By treating the CMNOs as regular MNOs, the same approach can be used to find the optimum weights except that the number of MNOs becomes larger.

The complexity order of the proposed scheduler is $O(BK)$. As seen in (4.7), RB b is assigned to the user who has the highest metric value $\alpha_m \eta_{k,b}$. Finding the maximum metric value among K users requires K operations. Consequently, assigning B RBs requires BK operations.

4.5.3 Computing the Optimal Weights for MNOs

Algorithm 4.2 presents an offline iterative search method that finds the optimal weights for all MNOs such that constraint (4.4c) is maintained. At each iteration, the mean squared error e is found between the agreed upon value r_m and the actual one Λ_m achieved using the current weights. The algorithm terminates either if the change in error Δe_{mse} falls below a threshold $e_{\min}$ or the maximum number of iterations $N_{\max}$ is reached. Two groups of MNOs are defined, G^+ which contains indices of MNOs with higher RAPs than the agreed upon values, and G^- which contains indices of MNOs that have RAPs less than the agreed upon values. At each iteration, δ is taken from the weights of MNOs that belong to G^+ and given to the weights of MNOs that belong to G^-. The weights of MNOs who belong to G^+/G^- decrease/increase proportional to their error e_m as seen in lines 13 and 15 of Table 4.2.

Table 4.2: Iterative search method to find MNOs' weights

while $\Delta e_{mse} < e_{\min}$ **and** $n < N_{\max}$ **do**
  $G^+ = \emptyset,\ G^- = \emptyset$
  for $m \in M$ **do**
    find $e_m^* = \Lambda_m^* - r_m$
    if $e_m^* > 0$ **then**
      $G^+ \leftarrow m$
    else
      $G^- \leftarrow m$
    end if
  end for
  for $m \in M$ **do**
    if $e_m > 0$ **then**
      $\alpha_m^* = \alpha_m + \delta\, e_m / \sum_{m \in G^+} e_m$
    else
      $\alpha_m^* = \alpha_m + \delta\, e_m / \sum_{m \in G^-} e_m$
    end if
  end for
  $e_{mse}^* = \frac{1}{M} \sum_m e_m^2$
end while

4.5.4 Numerical Results

In this section we study the performance of the proposed WRV scheme, with is denoted by "WRV LT". The performance of the proposed scheme is compared with two schemes. The first scheme is the static sharing (SS) scheme, which statically divides the spectrum into blocks and assigns them to the different MNOs. The second scheme is WRV short-term fairness scheme (denoted by "WRV ST"), which solves (4.4) for one time slot and allocates fixed number of RBs to each users each time slot. Three performance metrics are evaluated, the aggregate throughput, the RAPs of MNOs, and Jain's fairness index of the users of each MNO. The simulation assumes that there are 3 MNOs where the number of users for each MNO changes between 5 and 50. The sharing ratios are $r_1 = 20\%$, $r_2 = 35\%$, and $r_3 = 45\%$. Each user experiences a different average SNR in the range between 5 and 15 dB. The system bandwidth is assumed to be 20 MHz with 100 RBs available for allocation, and the error threshold is $e_{min} = 1 \times 10^{-4}$.

Fig. 4.4 shows that WRV LT scheme provides a higher aggregate throughput for all MNOs when compared to the SS scheme. The gain in aggregate throughput ranges between 11% and 17%. This is due to the fact that the SNRs experienced by users are considered in the scheduling decision rather than just statically assigning fixed blocks to each MNO. Moreover, the multiuser diversity with respect to spectral efficiency is also evident. For small number of users, the aggregated throughput improves as the number of users increases due to the multiuser diversity gain. However, for large number of users, multiuser diversity gain slightly increases with the number of users as the system reaches its the maximum capacity. In addition, WRV LT scheme outperforms WRV ST.

Fig. 4.5 shows the average user throughput for different MNOs. It can be noted from the figure that the average throughput per user decreases as the number of users increases. Such behavior is expected because there will be more competition for RBs. Therefore, users will receive fewer RBs and consequently have a lower average throughput.

Fig. 4.6 compares the analytical and simulated average RAP results. While the average SNR is changing, the proposed iterative scheme remains able to maintain the agreed upon RAP for each MNO, thus satisfying their users' requirements. The average percentage errors between the analytical and simulation results of the RAPs for different SNR

Table 4.3: Number of iterations to converge

N_{max}	10	20	30	50
$\delta = 2 \times 10^{-}3$	10	20	30	42
$\delta = 8 \times 10^{-}3$	10	13	13	13
$\delta = 16 \times 10^{-}3$	8	8	8	8
$\delta = 20 \times 10^{-}3$	10	20	30	50

values are less than 2%, which confirms the efficiency of the proposed scheme to satisfy the service level agreement for all MNOs as the average simulation RAPs are almost equal to the analytical results.

Simulation results indicate that the proposed scheme maintains a Jain's fairness index of ≥ 0.96. Jain's fairness index is calculated for the users of MNO m as

$$J(x_1, x_2, ..., x_{K_m}) = \frac{\left(\sum_{i=1}^{K_m} x_i\right)^2}{K_m \sum_{i=1}^{K_m} x_i^2} \tag{4.17}$$

where x_i is the rate of user i of MNO m.

Table 4.3 shows the number of iterations needed to converge for different step sizes and values of N_{max}. For $\delta = 2 \times 10^{-3}$, the algorithm takes 42 iterations to converge. For larger step sizes of 8×10^{-3} and 1.6×10^{-2}, the convergence rate is significantly reduced to 13 and 8 iterations respectively. However, N_{max} iterations are executed for $\delta = 20 \times 10^{-3}$ because the algorithm overshoots the optimum weights. Therefore, an optimal step size is needed to minimize the number of iterations for convergence.

4.6 Chapter Summary

This chapter demonstrates the concepts and benefits of the WRV. In addition, scope and depth of WRV are tackled with different solutions. The obtained results revealed that deep resource sharing enhances the resources utilization efficiency. However, the network becomes more complex and the isolation between the operators sharing the same network decreases.

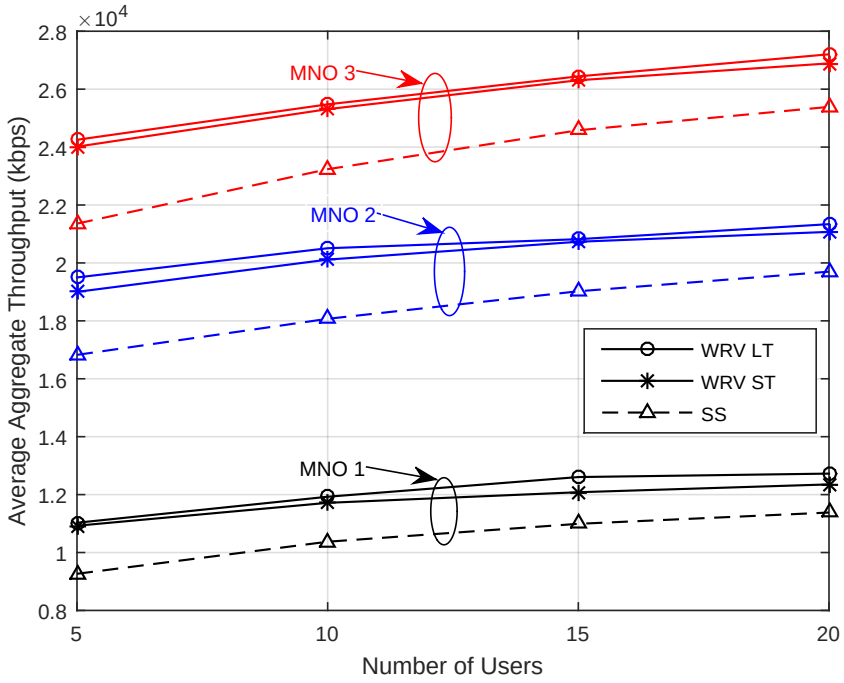

Figure 4.4: The average aggregate throughput of MNOs as the number of users increases.

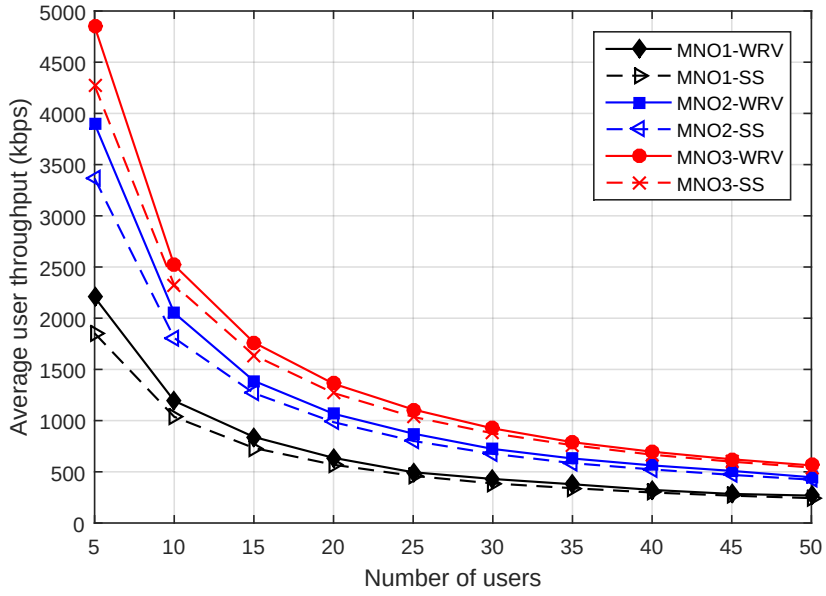

Figure 4.5: The average users' throughput of MNOs as the number of users increases.

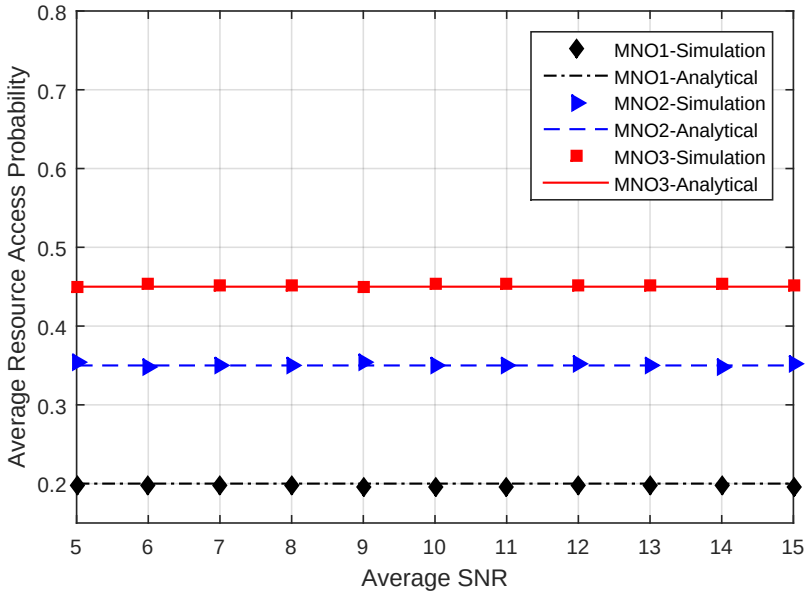

Figure 4.6: Resource access probability analytical and simulation results.

Furthermore, a framework for wireless resource virtualization under users' and MNOs APF constraint on a single eNB is presented. Normalized SNR is used to maintain fairness between users of a single MNO. To ensure fairness between different MNOs, the normalized SNRs are optimally weighted. An iterative search method is applied offline to determine the optimal weights. The iterative scheme has low-complexity and scales linearly with the number of RBs and users. Simulation results show that the proposed WRV scheme can improve throughput and maintain fairness between users and MNOs.

Chapter 5
Wireless Resources Virtualization for Cloud Radio Access Networks (C-RAN)

5.1 Introduction

Mobile network operators (MNOs) are highly interested in devolving cost-effective, scalable, and innovative solutions in order to improve network capacity and coverage to handle the ever-growing demand for mobile data transmission. Traditional solutions of deploying more base stations (BSs) and acquiring new spectrum licenses are becoming economically unsustainable. Therefore, cellular wireless networks need new techniques to enhance the utilization of computing and wireless resources. Recent spectrum utilization measurements show that the bandwidth licensed to MNOs is highly underutilized [73, 74]. According to a recent study by Nokia [7], only 20% of the radio access network's maximum capacity is used at any given time while 80% is being idle and waiting for peak hour demand.

Typically, the average traffic load in cellular networks is considerably lower than the peak load. In addition, the peak load time depends on the geographical location, because different locations might have different peak load times. Conventionally, BSs consist of network elements running purpose-built software and dedicated hardware. Therefore, efficient network design becomes very challenging because the network must employ sophisticated scalability techniques to handle a wide range of traffic loads at different times. Resource over-provisioning is costly, while under-provisioning compromises service quality. Moreover, adding new services or modifying existing ones is not a trivial process because it creates an impediment to network scalability, increases the time to market, and obstructs efficient utilization of the infrastructure.

With the growth of cloud-hosted development platforms and services, new promising solutions are on the horizon for MNOs. In 2010, China Mobile introduced the Cloud Radio

Access Network (C-RAN) as a new platform for future mobile network infrastructure [75]. In C-RAN, the baseband processing resources are aggregated and virtualized in a cloud data center, which allows better utilization of computing resources and more cooperative sharing of radio resources across wireless nodes. C-RAN technology promises MNOs significant reductions in capital expenditure (CAPEX) and operating expenditure (OPEX), as well as reduced energy consumption. Using C-RAN is expected to reduce power consumption by about 71% compared to traditional RAN [76]. However, to deployment of C-RAN commercially, many challenges should be addressed such that the transport network needs to be economical and supports high bandwidth, strict latency and jitter. More information about C-RAN deployment challenges can be found in [76].

The main concept of C-RAN architecture is to decouple the radio function and processing units of the RAN, as shown in Fig. 5.1, where a low-cost remote radio head (RRH) replaces the radio function unit. The baseband processing functions and resource management are delegated to baseband processing units (BPUs), which are pooled on a cloud in remote data centers. In traditional RAN architectures, the processing resources of BSs cannot be shared or scaled based on the traffic load. Since the processing power in C-RAN is centralized, innovative cloud-based solutions can be applied to improve the utilization of processing resources. Consequently, C-RAN architecture requires fewer BPUs compared to the traditional RAN architecture [77, 78].

Centralizing baseband processing and management enables better coordination across RRHs because cell site information, such as traffic loads, user-channel conditions, and user-traffic requirements, are available across the network. Such information can be effectively exploited to optimize the allocation of radio resources across cell sites, manage intercell interference (ICI), and improve coverage and handover procedures. In addition, sharing information can enhance capacity by facilitating the implementation of new technologies such as ICI coordination (ICIC), self-organizing networks (SON), and coordinated multi-point (CoMP) transmission.

Similar to C-RAN, which centralizes the computing resources, wireless resource virtualization (WRV) is a new paradigm aims at centralizing wireless spectrum resources and sharing them between MNOs. WRV enables network operators to create multiple logical networks on a single physical substrate yielding better efficiency in terms of energy

consumption and resource utilization. A new study from ABI Research [79, 80] shows that over a period of five years, deploying active infrastructure-sharing worldwide can save up to 60 billion USD in OPEX and CAPEX.

As MNOs need to continually enhance the capacity of their mobile networks, sharing radio resources between multiple MNOs facilitates carrier resource aggregation and supports higher peak rates. It also introduces multi-MNO multiplexing gains as a result of increasing the number of users per cell. For instance, in Rayleigh fading channels, the aggregated capacity of a cell can increase by $\ln(K)$, where K is the number of users in the cell [81]. Furthermore, network sharing facilitates new business models in the wireless market. For example, operators without Long Term Evolution (LTE) licenses or network resources will be able to provide LTE services by renting LTE radio resources from other MNOs.

Combining C-RAN and WRV should help MNOs to overcome the challenges inherent in current mobile networks. Therefore, the main objective of this chapter is to propose a new model that virtualizes the spectrum resources of cloud-based RANs, then efficiently share them between multiple MNOs' users. An entity called hypervisor is added on top of the physical network as shown in Fig. 5.1, which is responsible for allocating the available wireless resources to users subscribed to different MNOs. The allocation of wireless resources is determined by the sharing contract between MNOs, the traffic load at each RRH, and the interference between different RRHs.

The rest of this chapter is organized as follows: Section 5.2 provides a comprehensive literature review. Section 5.3 presents the system model, and problem formulation. The optimal solution of the proposed model is presented in Section 5.4. Suboptimal low-complexity schedulers are presented in Section 5.5. Section 5.6 presents and discusses simulation results, and Section 5.7 concludes the chapter.

5.2 Related work

In the recent literature, it can be clearly noticed that there is a growing interest in virtualizing networks' wireless resources [80]. For example, a wireless resource virtualization scheme for LTE evolved Node-B (eNB) is investigated in [62]. The scheme allows MNOs

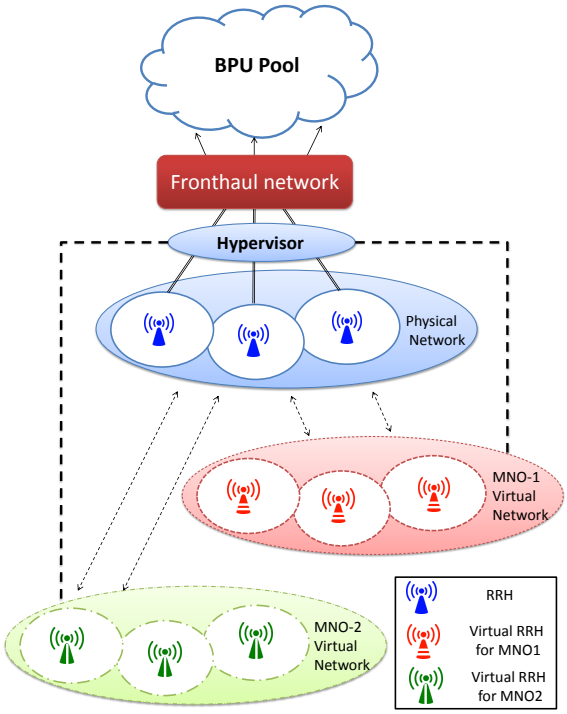

Figure 5.1: Virtualized C-RAN shared between two MNOs.

to implement different scheduling policies. However, the scheme did not consider network-wide virtualization nor coordination between interfering cell zones to prevent ICI. In addition, the scheme suffers from high complexity because it requires solving two optimization problems to maintain isolation between MNOs. In the first optimization process, the resources of each MNO are allocated to their users. The allocation results of problem one are then fed to the second problem as constraints such that the throughput each user obtains is equal or greater than the throughput achieved when sharing is not considered.

An LTE air interface virtualization scheme is proposed in [67], where a hypervisor is added on top of the physical resources. The hypervisor is responsible for virtualizing the eNB into a number of virtual eNBs that can be used by different MNOs. It is shown that more capacity can be achieved by sharing spectrum resources between different MNOs. However, the scheme does not provide optimal solutions nor manage ICI. Furthermore, the instantaneous channel quality of users is not considered in the scheduling decisions, which limits multiuser diversity gain.

More practical scenarios that consider load balancing are studied in [68, 82], where

the hypervisor manages the sharing process of multiple eNBs among multiple MNOs. Nevertheless, only fixed resource allocation across BSs is considered. The load is balanced between multiple BSs by moving users from high-traffic cells to low-traffic cells. However, transferring users across cells increases handover overhead, and may degrade the system capacity since users may be transferred to BSs further away, which would reduce the quality of the wireless link.

Another framework for wireless network virtualization that separates service providers from a network operator is reported in [83]. The service providers (SPs) are responsible for QoS management, while the network operator is responsible for spectrum management. The interaction between SPs and the network operator is modeled as a stochastic game regulated by the network operator. The role of the SPs is to compete for wireless resources for each subscribed user.

Utility-based resource provisioning scheme for WRV with massive MIMO is investigated in [84]. A single BS equipped with a large number of antennas serves users of different service providers. The problem is formulated as a combinatorial optimization problem of high computational complexity. Consequently, a low-complexity solution for the combinatorial problem is derived by linear programming relaxation.

In multi-cell systems, the same frequency bands can be assigned to users in different cells, which is referred to as the frequency-reuse (FR) principle, which is used to increase both coverage and capacity. However, to minimize ICI, cells that use the same frequency bands should be separated by a sufficient distance. Several ICIC techniques have been proposed for multi-cell systems as described in [85] and the references listed therein. The most promising is the fractional FR (FFR), which is adopted by 3GPP LTE [86].

The performance of FFR has been extensively studied for traditional cellular networks [52, 87, 88]. For C-RAN architecture, a dynamic FR scheme based on FFR is proposed in [89]. The wireless resources are assigned to cell zones using a graph-coloring approach. Each color represents a certain segment of bandwidth. To minimize ICI, different colors should be assigned to interfering zones.

A dynamic interference coordination scheme for downlink multi-cell systems is presented in [71]. The allocation problem is divided into two sub-problems, one at the BS level and the other at the central controller. It is assumed that BSs are able to communicate

with each other using an X2 interface. At the BS level, each sector potentially allocates bandwidth chunks to its connected users. Then, each sector sends a request to the central controller. The request specifies a list of bandwidth chunks to be restricted at the dominant interfering zones. Conflicting requests are resolved by the central controller, which sends a refined list of chunks that should be restricted to each sector.

Minimizing network power consumption of C-RAN is investigated in [90], where the power consumption of the transport network and RRHs is considered. The authors assume that transport links and RRHs can support sleep mode. The problem is formulated as a joint RRH selection and power minimization beamforming problem. The network power consumption is reduced by minimizing the number of active RRHs and reducing their transmit power subject to QoS constraints. Through simulations, the authors show that the network power consumption can be notably reduced. The performance of CoMP transmission schemes in a C-RAN architecture for LTE-A Heterogeneous networks is studied in [91]. With C-RAN architecture, a larger number of RRH can be considered in CoMP transmission, which improves the transmission performance.

5.3 System and Sharing Models

Consider the downlink of a cloud-based RAN architecture shared between M MNOs, where N RRHs are distributed to cover a certain geographical area. The RRHs are connected to a pool of BPUs in remote data centers via transport networks such as optical transport networks. It is assumed that the wireless link quality and expected traffic load are known for each user at the data center, which is responsible for resource allocation decisions. It is assumed that orthogonal frequency-division multiple access (OFDMA) is used for the downlink transmission. The total number of RBs available in the network is R. Each RRH is assumed to be capable of transmitting over any RB. The total number of user equipments (UEs) served by the network is K; each UE communicates with a single RRH.

Without loss of generality, it is assumed that each UE connects to the nearest RRH, and is labeled by a unique index $k \in [1, 2, \cdots, K]$. A table that maps each UE to a

particular MNO and an RRH is assumed to be available for BPUs. To facilitate readability, Table 5.1 summarizes the notations frequently used throughout the chapter.

Table 5.1: Summary of the most significant notation.

Symbol	Meaning
K, k	Total number of UEs, UE index
R, r	Total number of RBs, RB index
N, n	Total number of RRHs, RRH index
M, m	Total number of MNOs, MNO index
$P_{r,n}$	Transmit power from RRH n over RB r
$\mathcal{R}_m$	Set of RBs assigned to MNO m
R_m	Number of RBs assigned to MNO m
$\mathcal{R}_{m,n}$	Set of RBs assigned to MNO m at RRH n
$R_{m,n}$	Number of RBs assigned to MNO m at RRH n
$\mathcal{R}_n$	Set of all RBs assigned to RRH n
$\mathcal{K}_n$	Set of all UEs that connect to RRH n
$\mathcal{K}_{m,n}$	Set of UEs that subscribe to MNO m at RRH n
$K_{m,n}$	Number of UEs that subscribe to MNO m at RRH n
$\gamma_{r,k}$	SINR of UE k over RB r
$H_{r,n,k}$	Channel gain of the link between UE k and RRH n over RB r
$\mathcal{C}_n$	Set of RRH that interfere with RRH n
$\Phi_{m,n}$	Service status of MNO m at RRH n
Φ_m^{th}	Service status threshold of MNO m
β	Binary number indicator
$\Lambda_{m,n}$	Number of RBs allocated to MNO m at RRH n

The number of bits that can be transmitted over one RB is determined by the signal-to-interference plus noise ratio (SINR) at the receiver, which depends on the transmit power, channel fading parameters, interference introduced by other RRHs, and thermal noise. Therefore, the SINR of the link between UE k and RRH n over RB r can be expressed as

$$\gamma_{r,k} = \frac{P_{r,n} H_{r,n,k}}{\sum\limits_{c \in \mathcal{C}_n} P_{r,c} H_{r,c,k} + \sigma^2} \tag{5.1}$$

where $P_{r,n}$ is the transmit power of RRH n over RB r, $H_{r,n,k}$ is the channel gain of the wireless link between RRH n and UE k over RB r, which includes antenna gain and directivity, path loss, small-scale fading, and shadowing, σ^2 is the additive white Gaussian noise (AWGN) variance, and $\mathcal{C}_n$ is the set of RRHs that interfere with RRH n.

Although it is proven that power allocation schemes such as the water-filling algorithm can improve the transmission efficiency, integrating such techniques in mobile systems is challenging because they require tight tracking of the rapid channel variations [92]. Moreover, it is shown that, equal power allocation is as efficient as the optimal power allocation when the transmission power is high [93]. In this chapter, it is assumed that equal power allocation similar to the 3GPP-LTE standard [47, 94] such that $P_{u,n} = P_{v,n} = P_n, \forall\{u, v\} \in \mathcal{R}_n$, where $\mathcal{R}_n$ is the set of RBs assigned to RRH n. In fact, equal power allocation is a common assumption in resource allocation problems [71, 95, 96].

The indices of RRHs that interfere with each other can be described by an $N \times N$ binary symmetric matrix $\mathbf{C}$, where the element $c_{u,v} = 1$ if RRH u interferes with RRH v; otherwise, $c_{u,v} = 0$. RRH u is assumed to interfere with RRH v if the distance between RRH u and RRH v is less than a threshold.

In this chapter, the LTE physical layer model that supports various modulation and coding schemes (MCSs) [1] is considered. The maximum number of bits that can be transmitted to UE k over RB r can be calculated as

$$Q_{r,k} = \lfloor \xi(\gamma_{r,k}) T_{sym} \rfloor \tag{5.2}$$

where $\xi(\gamma_{r,k})$ is the spectrum efficiency of the selected MCS, T_{sym} is the total number of symbols in a single RB, and $\lfloor x \rfloor$ refers to the floor function.

As discussed in Section 2.3, the selection of a particular MCS is determined by using a lookup table which maps the received SINR to a MCS for a certain BLER as shown in Table 2.2.

5.3.1 Resource Blocks Sharing Model

RBs are assumed to be shared between MNOs based on a contract signed between them. RBs either shared statically, where each MNO accesses only its share of the resources, or dynamically, where MNOs can access the entire set of RBs. As wireless resource virtualization is still in its infancy stage, no well-defined sharing models exist yet [80]. Therefore, a general sharing model is assumed based on the following three conditions:

1. In the case of static sharing of RBs, MNOs are assumed to distribute their resources among cells such that the frequency reuse factor is maximized while maintaining a proportional fairness criterion such that

$$\max \sum_{n} \Lambda_{m,n} \tag{5.3a}$$

subject to

$$\Lambda_{m,n} + \sum_{c \in \mathcal{C}_n} \Lambda_{m,c} \leq \Psi_m, \forall n \tag{5.3b}$$

$$\Lambda_{m,1} : \cdots : \Lambda_{m,N_m} = (1 \pm \alpha)(L_{m,1} : \cdots : L_{m,N_m}) \tag{5.3c}$$

where $\Lambda_{m,n}$ is number of RBs allocated to MNO m at RRH n, $L_{m,n}$ is the load of MNO m at RRH n, and $\mathcal{N}_m$ is the set of RRHs that serves the UEs subscribed to MNO m. The load can be considered as the number of users or a number of packets queued in buffers for the users. As the fluctuation rate of the load in RRHs is slow compared with the transmission time interval (TTI), which is 1 ms in LTE systems for the finest scheduling granularity, the optimization problem in (5.3) can be solved at a coarser granularity than TTI. Other sharing models can be applied here, however, maximizing the frequency reuse factor while considering a fairness criterion is an intuitive target that MNOs are looking to achieve.

2. In case of dynamic sharing of the RBs, the service status of MNO m at RRH n should be higher than a certain threshold or, if it is not the case, MNO m should access at least $\Lambda_{m,n}$ RBs. This condition ensures isolation between MNOs such that all MNOs are either satisfied, or can access at least the same number of RBs that they would access in case of static sharing. The service status of an MNO can be related to aspects such as queue length of users' buffers, spectral efficiency, or energy efficiency.

5.4 Problem Formulation

It is assumed that each MNO aims at maximizing its sum weighted data rates, which is a very common optimization problem in wireless systems [11, 26, 71, 97]. The weights are

selected by MNOs according to their scheduling policies. Assume that user k is connected to RRH n, the scheduling problem can be formulated as

$$\max \sum_{m=1}^{M} \sum_{n} \left[\sum_{k \in \mathcal{K}_{m,n}} \sum_{r=1}^{R} \hat{w}_k u_{r,k} \, \beta_{r,k} \right] \tag{5.4a}$$

subject to

$$\sum_{c \in \mathcal{C}_n} \sum_{k \in \mathcal{K}_c} \beta_{r,k} + \sum_{k \in \mathcal{K}_n} \beta_{r,k} \leq 1, \forall n, r \tag{5.4b}$$

$$\sum_{r \in \mathcal{R}_k} T_{r,k} \leq q_k, \forall k \tag{5.4c}$$

$$(\Phi_{m,n} > \Phi_m^{th}) \text{ or } (\Psi_{m,n} \geq \Lambda_{m,n}) \text{ must hold, } \forall (m,n) \tag{5.4d}$$

where $u_{r,k}$ is the data rate achieved by assigning RB r to UE k, $\hat{w}_k$ is the normalized weight for UE k, $\mathcal{K}_n = \bigcup_m \mathcal{K}_{m,n}$ is the set of UEs connected to RRH n, $\mathcal{K}_{m,n}$ is the set of UEs subscribed to MNO m and connect to RRH n, $\Psi_{m,n}$ is the number of RBs accessed by MNO m at RRH n, R_k is the set of RBs assigned to UE k, $\Phi_{m,n}$ and Φ_m^{th} are the service status and service status threshold of MNO m at RRH n, and $\beta_{r,k}$ is a binary number indicator defined as

$$\beta_{r,k} = \begin{cases} 1, \text{ if RB } r \text{ is assigned to UE } k \\ 0, \text{ otherwise.} \end{cases}$$

Constraint (5.4b) represents the exclusive constraint which ensures that (i) each RB is assigned to one UE (at most) at each RRH, and (ii) orthogonal sets of RBs are allocated to RRHs that may interfere with each other. It is assumed that the interference is avoided if interfering RRHs are granted orthogonal sets of RBs. Constraint (5.4c) ensures that the transport block size for every UE is less than its unserved data size, where $\mathcal{R}_k$ is the RB set that is assigned to user k. Constraint (5.4d) specifies whether the service status of MNO m at RRH n is higher than a certain threshold or, if that is not the case, MNO m should access at least $\Lambda_{m,n}$ RBs. This constraint ensures isolation between MNOs such that MNOs are either satisfied, or can access at least the same number of RBs in case of static sharing. It is noteworthy that constraint (5.4d) can be split into two constraints by introducing a binary

variable $y_{m,n}$ and a sufficiently large upper bound B_m so that

$$\Phi_{m,n} > \Phi_m^{th} - B_m y_{m,n} \tag{5.5a}$$

$$\Psi_{m,n} \geq \Lambda_{m,n} - B_m(1 - y_{m,n}). \tag{5.5b}$$

When $y_{m,n} = 0$, constraint (5.5a) holds, whereas constraint (5.5b) becomes $\Psi_{m,n} \geq \Lambda_{m,n} - B_m$, which is always satisfied if B_m is large enough. Note that the constraint $\Psi_{m,n} \geq \Lambda_{m,n}$ may still be satisfied. When $y_{m,n} = 1$, only constraint (5.5b) holds. Consequently, one constraint holds, and the other one may be satisfied.

The formulation in (5.4) allows MNOs to apply different scheduling policies by weighting their UEs differently. In addition, it guarantees that MNOs use their share of RBs at the overloaded RRH. However, if an MNO is underloaded at a specific RRH, its share of RBs can be granted to other MNOs that are overloaded.

5.4.1 Special Case: Backlogged Traffic Model

In this subsection, a low-complexity formulation of (5.4) is presented for backlogged traffic model, where users always have data to transmit. Although such model might not occur all the time in the network, it reduces the complexity of the optimal solution. Therefore, the backlogged traffic model is assumed for optimal solution comparisons. Backlogged traffic model assumption relaxes the constraints (5.4c) because the data in users' buffers are assumed to be larger than the transmitted data. Therefore, if RB r is assigned to MNO m at RRH n, the RB should assigned to the user at RRH n who maximizes the weighted sum data rates and belongs to MNO m. Consequently, the scheduling problem for backlogged traffic model can be formulated as

$$\max \sum_{m=1}^{M} \sum_{n} \sum_{r=1}^{R} u_{r,m,n}^{max} \, \beta_{r,m,n} \tag{5.6a}$$

subject to

$$\sum_{c \in \mathcal{C}_n} \beta_{r,m,c} + \beta_{r,m,n} \leq 1, \forall r, n, m \tag{5.6b}$$

$$(\Phi_{m,n} > \Phi_m^{th}) \text{ or } (\Psi_{m,n} \geq \Lambda_{m,n}) \text{ must hold, } \forall(m,n) \tag{5.6c}$$

where $u_{r,m,n}^{max} = \max \sum_{k \in \mathcal{K}_{m,n}} \hat{w}_k u_{r,k}$ and $\beta_{r,m,c}$ is a binary number indicator defined as

$$\beta_{r,m,n} = \begin{cases} 1, \text{ if RB } r \text{ is assigned to MNO } m \text{ at RRH } n \\ 0, \text{ otherwise.} \end{cases}$$

5.4.2 Radio Resource Scheduling Polices

In wireless networks, radio resource scheduling plays a vital role in achieving maximum spectrum utilization, QoS satisfaction, and fairness between UEs. To achieve such goals, users are weighted according to the UEs service status and the scheduling policy. The service status of a user can be related to aspects such as queue length, traffic priority, and past performance levels achieved. The product of the weight and the data rate achieved using a specific RB interprets the users priority of using the RB. It is worth mentioning that the users' weights may vary with time.

Various scheduling policies are proposed for LTE networks [2], including channel-aware policies, such as Maximum Throughput (MT), Proportional Fair (PF), and Generalized PF (GPF); channel-aware and QoS-aware policies, such as Modified Largest Weighted Delay First (M-LWDF) and LOG rule; and energy-aware policies [8]. Table 5.2 illustrates examples of scheduling policies along with their types, targets, and weight definitions.

5.4.3 Complexity of Optimal Solution

The scheduling problem in (5.4) is a BIP optimization problem, which is NP-hard. The complexity of solving such optimization problems is considerably high and it increases exponentially with the number of users, MNOs, and RBs. Therefore, obtaining the optimal solution is computationally prohibitive even for a single MNO [71, 96].

In order to give a glimpse of the complexity of the BIP problem, one part of the problem is considered, which involves assigning RBs to RRHs. Consider a simple scenario where two MNOs share three RRHs that interfere with each other. Assume that the total number of RBs in the network is 60. As the RRHs interfere with each other, orthogonal

Table 5.2: Examples of different schedulers [2]

Scheduler	Type	Target	Weight
MT	channel-aware, QoS-unaware	throughput	$w_k = 1$
PF	channel-aware, QoS-unaware	fairness & throughput	$w_k = 1/\bar{T}_k$, where $\bar{T}_k$ is the past average throughput for UE k.
GPF	channel-aware, QoS-unaware	fairness & throughput	$w_k = (u_k)^{\eta_1}/(\bar{T}_k)^{\eta_2}$, where η_2, η_2 are constants tune the scheduler.
M-LWDF	channel-aware, QoS-aware	queue stability	$w_k = -D_k log(\alpha_k)/D_k^{max}$, where α_k is the acceptable delay violation probability, D_k^{max} is delay threshold, and D_k is the head-of-line packet delay.
LOG rule	channel-aware, QoS-aware	queue stability	$w_k = b_k log(c_k + a_k D_k)$, where a_k, b_k, and c_k are constants tune the scheduler.

sets of RBs should be assigned to RRHs. If each RRH receives 20 RBs, the number of combinations of 20 RBs chosen from 60 RBs for one MNO is

$$\frac{60!}{20!(60-20)!} = 4.1918 \times 10^{15}.$$

If the finest scheduling granularity is chosen, which is one subframe (1 ms) for LTE systems, then the optimization problem should be solved in less than 1 ms; which is practically infeasible. Therefore, deducing low-complexity algorithms is crucial for solving the problem using reasonable computational power.

5.5 Low-Complexity Solutions

The optimal solution in (5.4) is achieved by jointly allocating RBs to users subscribed to different MNOs across all the available RRHs. However, it is computationally expensive to take all RBs in the scheduling problem into consideration, as described in Section 5.4.3. A possible approach to reduce the complexity is to allocate RBs sequentially. In this section, two iterative low-complexity solutions are proposed, each of which allocates a single RB at each iteration. The basic concept is that, if RB r is assigned to RRH n, it should be assigned to a user who is subscribed to the least satisfied MNO (LSM). As the objective is to maximize weighted sum utility, the assigned RB to the RRH is granted to the user who

can maximize the sum-utility. To allocate RB r, the set of LSM at every RRH is denoted by $1 \times N$ vector $\mathbf{s}_r = [s_{1,r}, \cdots, s_{N,r}]$, where $s_{n,r}$ is the index of the LSM at RRH n. The vectors $\mathbf{z}_r = [z_1, \cdots, z_{N,r}]$ and $\mathbf{j}_r = [j_{1,r}, \cdots, j_{N,r}]$ denote the utilities and indices, respectively, of the UEs subscribed to $\mathbf{s}_r$ and who have maximum utility, where

$$z_{n,r} = \max_{k \in \mathcal{K}_{sn,n}} \hat{w}_k u_{k,r}$$

$$j_{n,r} = \arg \max_{k \in \mathcal{K}_{sn,n}} \hat{w}_k u_{k,r}$$

The optimization problem per RB can be seen as the maximum weighted independent set (MWIS) of a graph path. Each RRH represents a vertex; an edge (line) is drawn between two vertices if they interfere with each other. A graph can be described by the pair $G = (V, E)$, where the set V is the vertices of G, and the set E is the edges of G. The MWIS is the subset of vertices that has maximum weighted sum such that no two vertices are connected with an edge. Fig. 5.2 shows an example of a graph $G = (V, E)$ of five weighted vertices (RRHs), where $V = \{1, 2, 3, 4, 5\}$, and $E = \{(1, 4), (1, 5), (2, 4), (3, 5)\}$. The independent sets are = $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{1, 2\}$, $\{1, 3\}$, $\{1, 2, 3\}$, $\{2, 3\}$, $\{2, 5\}$, $\{2, 3, 4\}$, $\{2, 4, 5\}$, $\{3, 4\}$, $\{4, 5\}$. The MWIS is $\{1, 2, 3\}$.

For each RB, the MWIS I_S can be found by solving the following optimization problem

$$I_S = \max \sum_{n=1}^{N} v_n \, \beta_n \tag{5.7a}$$

subject to

$$\sum_{c \in \mathcal{C}_n} \beta_c \leq 1, \forall n \tag{5.7b}$$

where β_n is a binary variable, equal to one if $n \in I_S$ and zero otherwise, and v_n is the weight of vertex (RRH) n. In order to bias the scheduler towards allocating RBs in favour of highly loaded RRHs, the weights are chosen such that

$$v_n = \begin{cases} z_{n,r} \, (\Lambda_{m,n} - \Psi_{m,n}), \text{if } \Phi_{s_n,n} < \Phi_{s_n}^{th} \\ z_{n,r}, \text{ otherwise.} \end{cases}$$

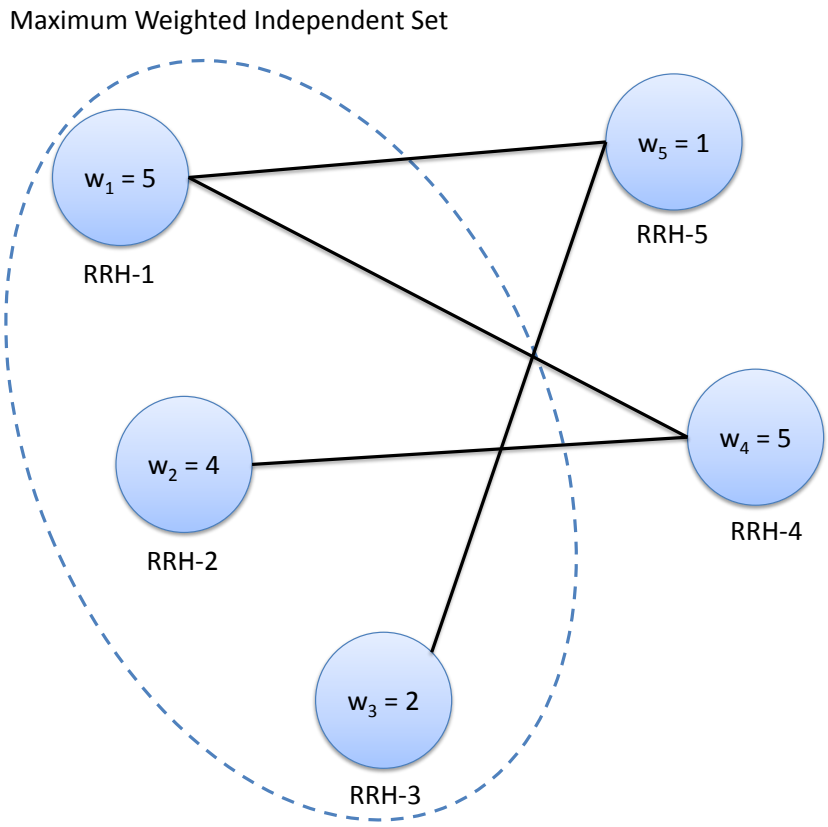

Figure 5.2: Example of interference graph $G = (V, E)$ of five weighted vertices (RRHs).

Table (5.3) shows the pseudo code of an iterative low-complexity algorithm that solves (5.7) by using a BIP solver. At each iteration, one RB is assigned to the MWIS of RRHs that maximizes the sum-weighted utility. Lines 5-8 find the LSM (s_n) at every RRH. To avoid assigning RBs to MNOs that have no data to transmit, the service status of the LSM has to be lower than the threshold $\Phi_{m,n} < \Phi_m^{th}$. Lines 9-10 find the index and the utility value of the user who subscribes to MNO s_n and maximizes the sum utility. In line 12, the algorithm solves the MWIS optimization problem (5.7) and finds the subset I_S. The RB is assigned to users in line 14. The number of RBs requited is decremented for each RRH that belongs to I_S in line 13, whereas $\bar{\Omega}_{s_n,n}, \Phi_{s_n,n}$ are updated in line 20. The algorithm runs until all RBs have been assigned.

Although the algorithm solves the BIP problem R times each TTI, the complexity of the algorithm is relatively low as compared to (5.4) because the size of the BIP optimiza-

Table 5.3: Per RB optimal allocation algorithm

1: **input:** $\Lambda_{m,n}, \bar{\Omega}_{m,n}, \forall m, n$
2: $\mathcal{R}_n = \mathcal{R}, \forall n$
3: **for** $r = 1 : R$ **do**
4: **for** $n = 1 : N$ **do**
5: $v_{n,r} = 0$
6: $\Delta\Omega_{m,n} = \Lambda_{m,n} - \bar{\Omega}_{m,n}, \forall m$
7: $\Omega_{m,n} = \Lambda_{m,n} + \Delta\Omega_{m,n}, \forall m$
8: $s_n = \arg\max_{m} \ \Omega_{m,n}, \forall m$ such that $\Phi_{m,n} < \Phi_m^{th}$
9: $z_{n,r} = \max_{k \in \mathcal{K}_{s_n,n}} u_{k,r}$
10: $j_{n,r} = \arg \max_{k \in \mathcal{K}_{s_n,n}} u_{k,r}$
11: **end for**
12: **solve** the BIP problem in (5.7)
13: **assign RB** r **to UE** $j_{n,r}, \forall n \in I_S$
14: $\Omega_{s_n,n} = \Omega_{s_n,n} - 1, \forall n \in I_S$
15: **update** $\bar{\Omega}_{s_n,n}$ & $\Phi_{s_n,n}, \forall n \in I_S$
16: **end for**

tion problem is significantly smaller that the problem in (5.4). In particular, the BIP has N decision variables. However, for a large number of RRHs, it might be computationally expensive to solve the BIP problem shown in (5.7) R times every TTI. Therefore, a low-complexity heuristic algorithm that solves the BIP is presented in Table 5.4. The heuristic algorithm is greedy in the sense that it assigns an RB to the RRH that has the maximum weight v_n, then excludes its interfering RRHs from the allocation process. The first 11 lines in the heuristic algorithm are similar to those in Table 5.3, where the LSMs and their users who maximize the sum utility are specified. The RRH index that has the maximum weighted utility n^* is found in line 14 and is added to the subset I_S in line 15. The RRH n^* and its interfering RRHs indices are deleted from the potential set of RRHs S_{ind}. Consequently, interfering RRHs are not assigned the same RBs, thereby eliminating interference. An RB is assigned to $j_{n,r}, \forall n \in I_S$ in line 18, and $\Omega_{s_n,n}, \bar{\Omega}_{s_n,n}, \Phi_{s_n,n}$ are updated in lines 19-20.

Table 5.4: Heuristic algorithm

input: $\Lambda_{m,n}, \bar{\Omega}_{m,n}, \forall m, n$
$\mathcal{R}_n = \mathcal{R}, \forall n$
for $r = 1 : R$ **do**
  for $n = 1 : N$ **do**
    $v_{n,r} = 0$
    $\Delta\Omega_{m,n} = \Lambda_{m,n} - \bar{\Omega}_{m,n}, \forall m$
    $\Omega_{m,n} = \Lambda_{m,n} + \Delta\Omega_{m,n}, \forall m$
    $s_n = \arg\max_{m} \ \Omega_{m,n}, \forall m$ such that $\Phi_{m,n} < \Phi_m^{th}$
    $z_{n,r} = \max_{k \in \mathcal{K}_{s_n,n}} u_{k,r}$
    $j_{n,r} = \arg \max_{k \in \mathcal{K}_{s_n,n}} u_{k,r}$
  end for
  $S_{ind} = \{1, 2, \cdots, N\}$
  while $S_{ind} \neq \phi$ **do**
    $n^* = \arg\max_{n} \ v_{n,r}$
    $I_S \leftarrow n^*$
    $S_{ind}^* = S_{ind} \setminus c \in \{\mathcal{C}_n \cup n^*\}$
  end while
  assign RB r **to UE** $j_{n,r}, \forall n \in I_S$
  $\Omega_{s_n,n} = \Omega_{s_n,n} - 1, \forall n \in I_S$
  update $\bar{\Omega}_{s_n,n}$ & $\Phi_{s_n,n}, \forall n \in I_S$
end for

5.5.1 The Complexity of the Heuristic Algorithm

For every TTI, the heuristic algorithm runs R major iterations (lines 3-21). Each major iteration finds the LSM and a candidate user $j_{n,r}$ for each RRH. Finding the LSM at RRH requires M operations, whereas finding the user ($j_{n,r}$) requires $K_{s_n,n}$ operations. The number of UEs is usually much larger than the number of MNOs, which makes finding $j_{n,r}$ is the dominating operation. Assigning each RB to the subset I_S requires at most N operations, assuming that no RRHs interfere with each other. Therefore, the worst-case complexity is $O(R \times (N + K^{max}))$, where $K^{max} = \max_{n,m} K_{m,n}$ is the maximum number of users that connect to an RRH and subscribe to one MNO.

5.6 Simulation Model and Numerical Results

In this chapter, a layout that comprises of 22 hexagonal cells as shown in Fig. 5.3 is considered. The number of users for each MNO at each cell is assumed to be uniform random variable with a mean of $\bar{K}$. UEs are assumed to be uniformly distributed across each cell and have average SINR of 5 to 10 dB. All channels are assumed to be independent and frequency-flat with Rayleigh fading. Each channel is assumed to be fixed during one subframe in time domain and over one RB in frequency domain, but changes independently over different subframes, different RBs, and different users which corresponds to a quasi-static channels.

For benchmarking purposes, the proposed schemes are compared to static sharing scheme and the heuristic algorithm presented in [62], which is which is referred to as heuristic-RRH. In the static sharing, each MNO at each RRH receives its share of RBs ($\Lambda_{m,n}$) and allocates them to its users according to the MNO scheduling policy. In the heuristic-RRH scheme, the spectrum resources assigned to all MNOs at a single RRH are virtualized and shared between all UEs connected to the RRH.

In the following two subsections, numerical results are presented for two different scenarios. The first scenario considers the optimal solution, and thus only RRHs 1-6 are considered and the backlogged traffic model is assumed. In the second scenario, the entire layout of the 22 cells is considered and dynamic traffic model is assumed. Therefore,

the optimal solution is excluded due to its high complexity while the proposed BIP and heuristic schemes are considered and compared to static sharing and the heuristic scheme presented in [62].

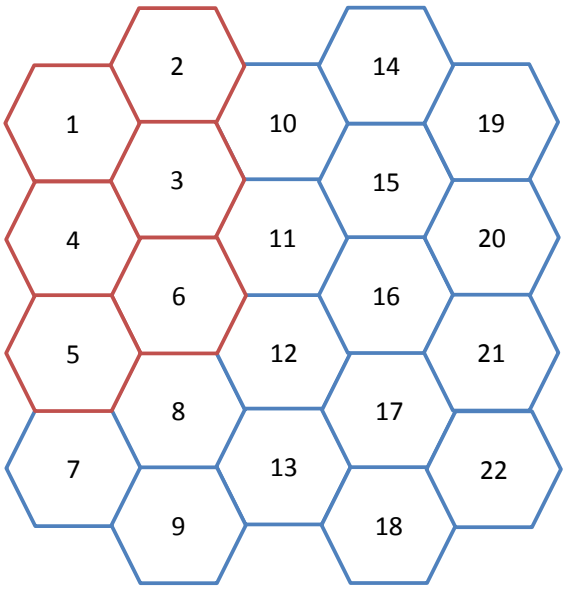

Figure 5.3: Simulated network layout. For the first scenraio, only the red RRHs 1-6 are considered. For the second scenario, all RRHs are considered.

5.6.1 Scenario 1: Backlogged Traffic Model

In this case, throughput of the limited-size layout is evaluated and compared for various schemes. The MT scheduling policy and the backlogged traffic model are adopted for both MNOs. The backlogged traffic model implies that each MNO is fully loaded and operates at full capacity. Therefore, the sharing gain is minimal and it is only due to the multiplexing gain of the resources. To demonstrate the benefit of sharing the resources between MNOs, users of MNO2 may hibernate for a random period of time with an average value of T_{sleep}.

Fig. 5.4 compares the average throughput per UE for different average number of UEs per cell per MNO. As it can be noted from the figure, the general trend for all schemes is that increasing the average number of UEs per cell per MNO decreases the average number of RBs assigned to each UE, which decreases the average throughput per UE. However,

each scheme provides its own average throughput based on the adopted resource allocation scheme. As expected, all the results are upper bounded by the optimal scheme and lower bounded by the static sharing scheme. Moreover, the proposed BIP and Heuristic schemes outperform the heuristic-RRH scheme and they are only 9% less than the optimal. The throughput advantage of the proposed BIP and heuristic approaches is due to the fact that UEs in [62] may access only an RBs that is assigned to their RRH. However, the heuristic-RRH scheme outperforms the performance of the static sharing scheme as it virtualizes the resources of both MNOs at each RRH. Static sharing offers the lowest throughput because the EUs of each MNO can access their dedicated RBs. And hence, UEs loose the chance to use underutilized RBs that belong to other MNOs or other RRHs. Although MNO1 and MNO2 own same number of RBs, the average throughput per user is lower for MNO2's users because they were forced to hibernate T_{sleep} of the time. The throughput of the considered schemes for MNO2's UEs follows their performance for MNO1's UEs, where the optimal scheme outperforms all other schemes. The average throughput per user for the proposed BIP and heuristic are slightly lower than the average throughput per user achieved by the optimal scheme. The heuristic-RRH scheme slightly outperforms the static sharing scheme. However, the gain achieved by the WRV schemes for MNO2's UEs is less than that for MNO1's UEs for the following reason. WRV offers two type of gains: 1) multi-MNO multiplexing gains as a result of increasing the number of users per cell, and 2) gain results form sharing RBs of underloaded MNOs with overload MNOs. As MNO1 is fully loaded, WRV only offers multi-MNO multiplexing gains to MNO2's users. On the other hand, MNO2 is assumed to be underloaded and MNO1's users benefit from both gains.

Fig. 5.5 shows the average aggregate throughput per cell for the five schemes. As number of UEs increases, the average aggregate throughput increases as a result of the multi-MNO multiplexing gain.

The average user throughput for different values of T_{sleep} for users of MNO1 and MNO2 are shown in Figs. 5.6 and 5.7, respectively. It is worth noting that T_{sleep} indicates the load of MNO2. As value of T_{sleep} becomes longer, the load on MNO2 becomes lighter. In case of static sharing, MNO1 and MNO2 are fully isolated from each other. Therefore, lightening the load on MNO2 has no impact on MNO1, and the average throughput of MNO1 users is constant for any T_{sleep} value. On the contrary, for WRV schemes the av-

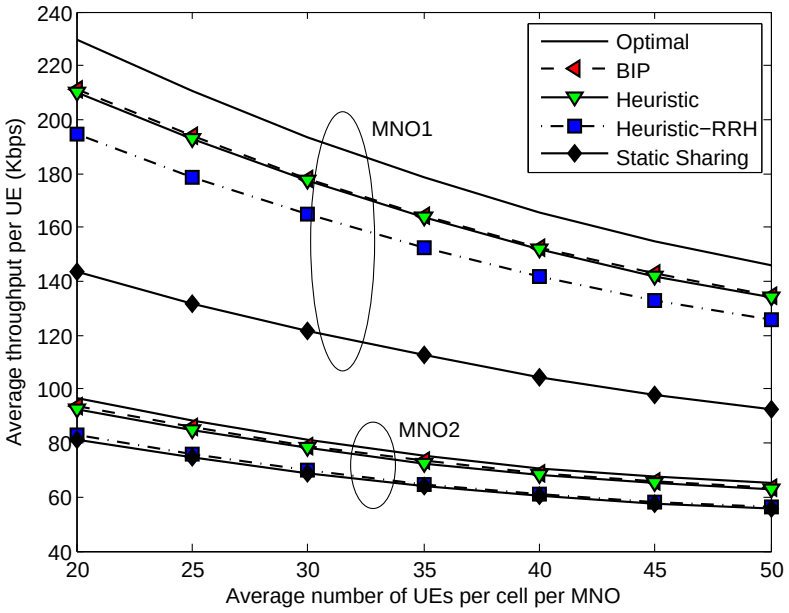

Figure 5.4: Average throughput per UE for $T_{sleep} = 40\%$.

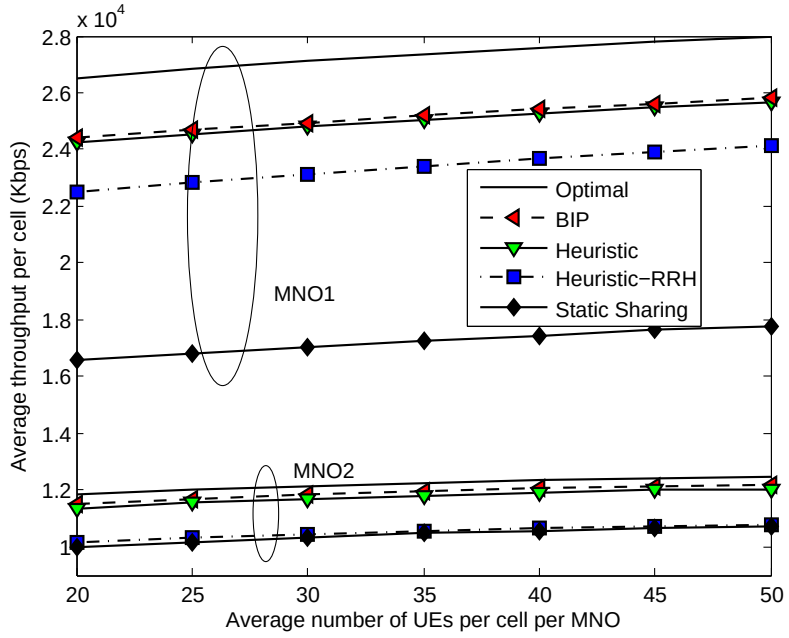

Figure 5.5: Average aggregate throughput per cell for $T_{sleep} = 40\%$.

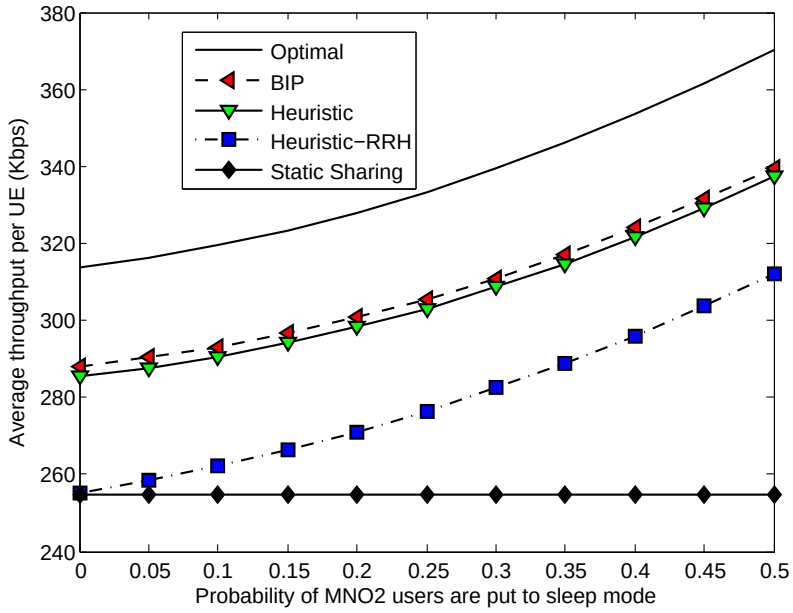

Figure 5.6: Average throughput per UE for MNO1's users for different vales of T_{sleep}.

erage throughput of MNO1's users build up as the MNO2 load drops. As T_{sleep} increases, the users of MNO2 access less RBs, and therefore their average throughput decreases or the benefit of MNO1's users.

5.6.2 Scenario 2: Dynamic Traffic Model

In this scenario, the average throughput and head-of-line (HoL) packet delay are evaluated. The simulation variables are similar to the first scenario in terms of the number of MNOs, and RBs. The assumptions and the system model of this scenario are more realistic because the entire layout which consists of 22 RRHs is considered. Moreover, MNOs may apply different scheduling polices. In this chapter, it is assumed that MNO1 applies the M-LWDF scheduling policy, while MNO2 applies the MT scheduling policy. UEs are active if they have data to transmit. The data traffic for UEs subscribed to MNO1 are modelled by Poisson traffic model with an average packet arrival rate λ [34, 71, 97], and fixed-size packet of 1 KB. UEs subscribed to MNO2 are assumed to be greedy and always have data to transmit.

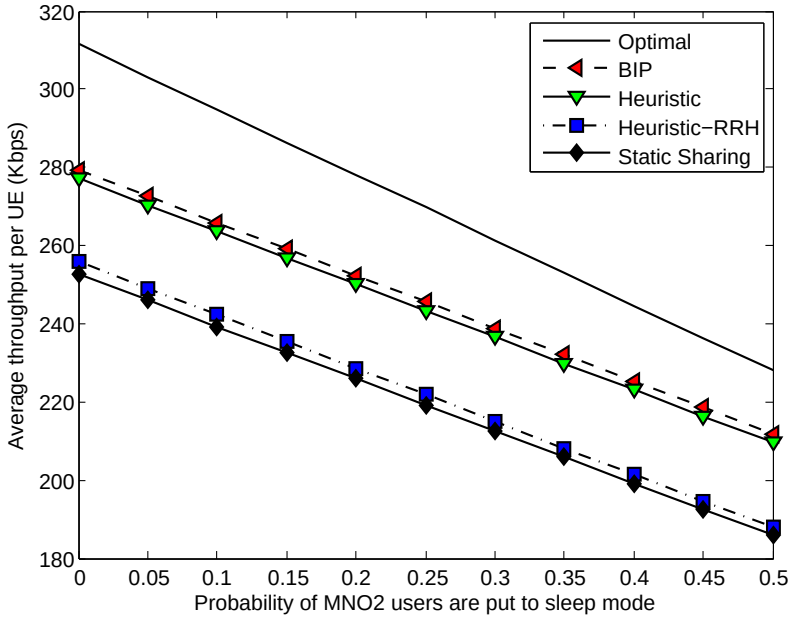

Figure 5.7: Average throughput per UE for MNO2's users for different vales of T_{sleep}.

The performance of the proposed, heuristic-RRH, and static sharing schemes is compared for different packet arrival rates λ, which is defined as the average rate at which packets arrive to the users' buffers. For example, $\lambda = 50$ implies that, on average, a 1 KB packet arrives every 50 ms. Since both MNOs have similar profiles in terms of the number of RBs, scheduling policy, number of UEs, and average SINR, an average result for both MNOs is shown.

The average aggregate throughput per cell for users of MNO1 and MNO2 is shown in Figs. 5.8 and 5.9, respectively. As the value of λ increases, the average data arrivals deceases. Consequently, MNO1 becomes further underloaded and the average cell throughput decreases. In the static sharing scenario, RBs assigned to MNO1 are not accessible by MNO2. Therefore, average aggregate throughput of MNO2 is not affected by the variation in the traffic load of MNO1. For WRV schemes, the throughput of MNO2 grows as the load of MNO1 becomes lighter. Reducing the load of MNO1 allows MNO2 to access more RBs that would be granted to MNO1 if it is overloaded. The BIP scheme slightly outperforms the proposed heuristic scheme. The performance of the proposed schemes is higher than the performance of the heuristic-RRH and static sharing schemes.

Fig. 5.10 shows the average HoL packet delay for different values of λ. The average HoL delay for the BIP and the heuristic schemes are similar, but much less than that for the heuristic-RRH and static sharing schemes for small value of λ. For large value of λ, the load on MNO2 is relatively low and average HoL delay for all schemes converge to low values.

The average running time of the schemes is measured and pointed out in Table 5.5. The static sharing scheme is the least complex solution, since the RB allocation is independent for each RRH. The heuristic solutions are solved considerably slower than static sharing but faster than the BIP scheme.

Table 5.5: Normalized average running time of the BIP, heuristic, heuristic-RRH and static sharing solutions.

BIP	Heuristic	heuristic-RRH	Static sharing
2.52×10^4	0.712×10^4	0.544×10^4	1

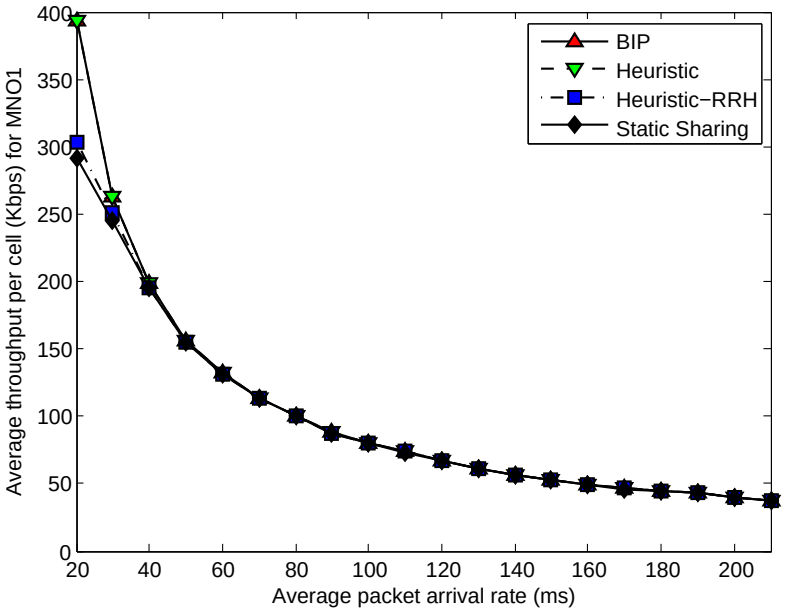

Figure 5.8: Average aggregate throughput of MNO1's users.

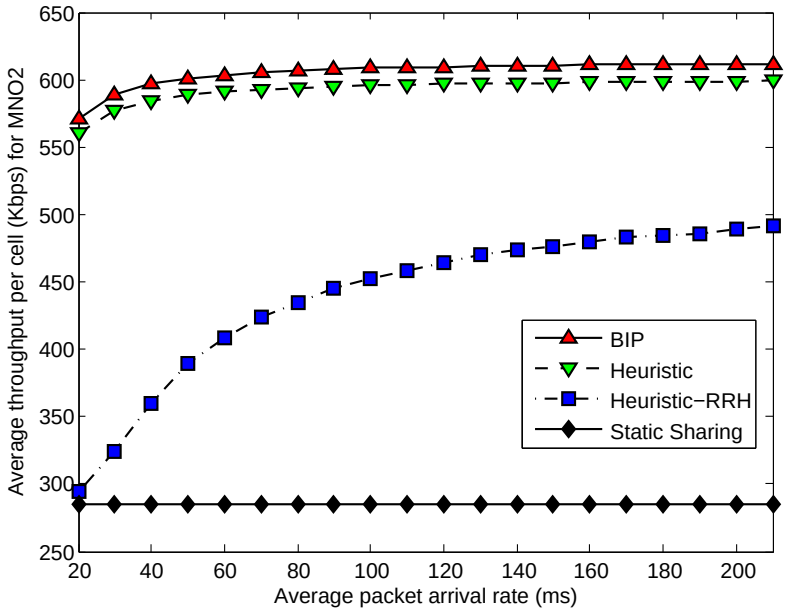

Figure 5.9: Average head-of-line packet delay of MNO2's users.

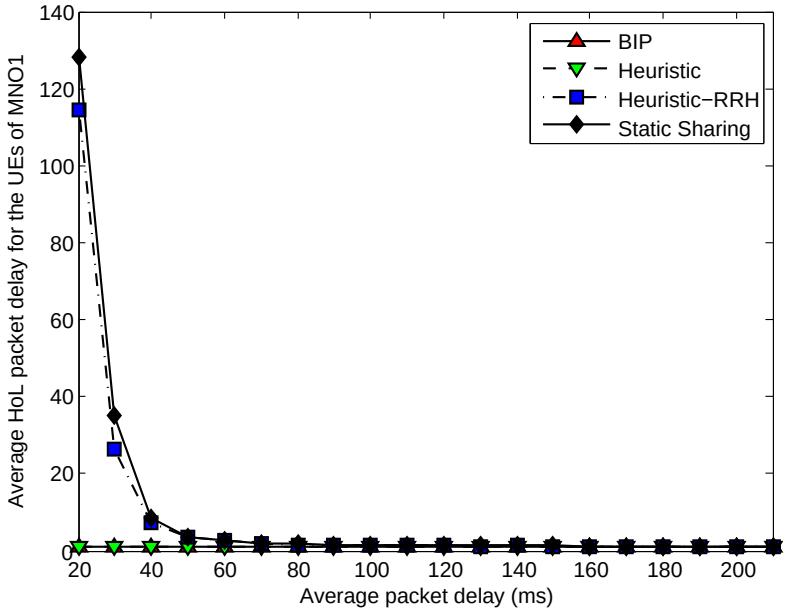

Figure 5.10: Average aggregate throughput of MNO1's users.

5.7 Chapter Summary

In this chapter, wireless resource virtualization schemes for cloud-based radio access networks (C-RANs) is presented. The proposed schemes dynamically share wireless resources between multiple mobile network operators (MNOs). Optimal and suboptimal solutions are provided and compared with each other as well as with a static sharing scheme, where each MNO is assigned a fixed set of radio RBs. The optimal solution is formulated as a binary integer programming optimization problem, which is known to be computationally expensive. Consequently, to reduce the complexity of the optimal formulation, two low-complexity suboptimal schemes are derived. The performance of the suboptimal solutions is slightly lower than the optimal solution at the benefit of a significant lower running time. The performance of the proposed schemes are compared in terms of throughput, delay, and time complexity. The simulation results show that the proposed schemes outperform static sharing in terms of both aggregate throughput and delay.

Chapter 6
Conclusion

The ever-increasing demand for high data rate transmission over mobile networks results in a new set of challenges for wireless systems and device designers such as improving spectrum efficiency and extending battery life-per-charge of the wireless devices. The main objective of this thesis is to provide efficient solutions for such challenges using optimization techniques. The thesis is divided into two parts. The first part is motivated by the fundamental requirements for extending time-per-charge utilization for battery-powered wireless devices. The second part of the thesis is motivated by the fundamental requirements for cost-efficient solutions to improve the wireless network capacity and maximize average revenue per user. In this section, the work presented in the thesis is summarized and recommendations for extending the research are proposed.

6.1 Thesis Summary

Chapter 2 and 3 focus on designing power-efficient packet schedulers that are able to minimize the total transmission power for users while maintaining QoS requirements. Two power-efficient schedulers for mixed streaming services in Long Term Evolution (LTE) uplink systems are presented in Chapter 2. The proposed schedulers are subject to rate, delay, contiguous allocation, and maximum transmission power constraints. We first derive an optimal scheduler that uses binary integer programming (BIP). Then, a low-complexity iterative scheduler that solves the BIP problem is presented. It is shown that the proposed schedulers maintained the required QoS and reduced the total transmit power under different practical scenarios. These power savings were achieved because of the schedulers transmit data at low rates while maintaining the required QoS.

For benchmarking purposes, we further find the global optimal solution using dynamic programming (DP), which requires the knowledge of future arrival rates and future channel gains for all users. Although the global optimal solution is non-causal and computationally expensive, it can be used as a reference for other sub-optimal algorithms. Analysing the optimal solution guides us to propose another two low complexity heuristic schedulers to solve the optimization problem. The first scheduler controls the maximum allowable transmit power (MATP) for each user based on the queue length. In particular, if the queue length of a user is relatively small, the scheduler reduces the users transmission rate as well as the MATP to minimize the total transmit power. On the other hand, when the queue length is large, the scheduler increases both the transmission rate and the MATP to satisfy the delay requirements. The second scheduler controls the minimum acceptable bit per Watt ratio (BPWR) for each user. Users can only transmit if their BPWRs are greater than an acceptable level, which allows only high power efficient transmission.

Chapter 4 and 5 present wireless resource virtualization (WRV) as a key technology to overcome the major challenges facing the mobile network operators (MNOs) such as reducing the capital, minimizing the operating expenses, improving the quality of service (QoS), and satisfying the growing demand for mobile services. In Chapter 4, the most dominant WRV frameworks are discussed, where different levels of network infrastructure and spectrum resources are shared between multiple MNOs. Moreover, the major benefits and most pressing business challenges of deploying WRV are summarized. Furthermore, an optimal low-complexity scheduler is proposed to virtualize the wireless resource blocks (RBs) and share them between users of MNOs. The scheduler aims at maximizing throughput while maintaining access proportional fairness among users as well as MNOs.

Solutions for virtualizing wireless resources of C-RANs and sharing them between multiple MNOs are presented in Chapter 5. The proposed solutions dynamically allocate wireless resources to users who subscribe to MNOs across the network. In addition, the proposed solutions maintain a high level of isolation between different MNOs, provide efficient resource utilization, enable different scheduling polices, and manage intercell interference.

Overall, the primary argument made in this thesis is that the proposed resource allocation and virtualization techniques can serve as basis for effective solutions to mitigate

the modern challenges faced by wireless system designers and users pertaining to energy efficiency, spectrum scarcity, and resources utilization.

6.2 Future Work

As the demand for mobile access and services is expected to rapidly grow [4], wireless network technologies will continue to develop, and become more sophisticated. Evidently there is much work to be done in the area of resource allocation and virtualization in next generation wireless networks (NGNs). The most direct extension of this work is to consider future technologies and network architectures while allocating and virtualizing wireless resources. This section proposes a variety of research directions for future work.

6.2.1 Power-Efficient Schedulers for LTE-A Uplink

The International Mobile Telecommunications-Advanced (IMT-Advanced) requirements motivate further development of LTE towards LTE-Advanced (LTE-A). IMT-Advanced requirements expect a data rate of 1Gbps in downlink and 500Mbps in uplink. New radio features specified in LTE-A such as multi-user multiple-input and multiple-output (MU-MIMO), Coordinated Multipoint (CoMP), and Carrier Aggregation (CA) aim to meet the IMT-Advanced requirements, and further improve cell coverage, throughput, and system efficiency. However, deploying such features adds degrees of freedom and extended dimensions to the resource allocation problems. Scheduling and resource allocation for high dimensional spaces is challenging (known as the curse of dimensionality). In this section, two LTE-A radio features are discussed from a scheduling point of view as a logical extensions of the previous works.

6.2.1.1 MU-MIMO

MU-MIMO is one of the key enablers for achieving high spectral efficiency in LTE-A. MU-MIMO achieves additional multiuser diversity gain by grouping users together such that they simultaneously transmit over the same time-frequency resources. Each group of users can be seen as one user with multiple antennas. However, the performance of

MU-MIMO depends closely on the method of grouping users. Designing power-efficient schedulers that jointly allocate wireless resources and group users in MU-MIMO systems can be a potential topic for future research in this area.

6.2.1.2 Carrier Aggregation

LTE-Advanced introduces the CA technology to meet the IMT Advanced requirements. With the CA feature, it is possible to aggregate two or more component carriers (CCs) together. The aggregated CCs have wider transmission bandwidth (up to 100 MHz) and convey higher data rates. From scheduling point of view, CA enables users to share resource over multiple CCs. Uplink scheduling with CA should consider the following issues:

1. How to assign and share CCs between users, which performs a critical role of the system performance, considering that different users have different aggregation abilities, depending on the EU type (LTE or LTE-Advance).
2. CC load balancing is needed for optimal system performance.
3. Users who transmit over multiple CCs might break the contiguous allocation constraints of SC-FDMA, and increase PAPR. One way to solve the issue of increased PAPR is to reduce the maximum transmission power, so the power amplifier operates in the linear region. On the other hand, transmission power reduction limits the users' throughput. The trade off between PAPR and users' throughput should be considered in the schedule design.

6.2.2 Virtualization in Next Generation Radio Access Network

A key enabler of the NGNs such as 5G is the integration of multiple radio access technologies (RATs) including 4G, 3G (UMTS/HSPA), GPRS/EDGE and Wi-Fi. Multi-RAT network performance can be improved by smartly utilizing the wireless resources available among each RAT [98]. Moreover, exploiting heterogeneous networks (HetNets) has emerged as a new network planning drift and as a promising solution to satisfy the growing demand for broadband wireless access networks. HetNets consist of a mix of technologies, frequencies, cell sizes, and network architectures to enhance capacity and coverage of

wireless networks. Chapter 4 and 5 investigate virtualizing wireless RANs that apply single radio access technology and single network architecture. Future work, therefore, should focus on providing solutions that are able to abstract and share the spectrum resources in NGN.

6.2.3 Spectrum and Computing Resources Virtualization in C-RAN

Chapter 5 focuses on virtualizing only the spectrum resources in cloud RANs (C-RANs). However, C-RAN architecture allows computing resources virtualization, where baseband processing units (BPUs) can be dynamically shared among remote radio heads (RRHs), yielding better computing resource utilization and power efficiency. Virtualizing both the spectrum and the compute resource between multiple operators is another interesting direction to extend this work.

References

[1] 3GPP.TS.36.213 11.0.0, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," 2012.

[2] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 2, pp. 678–700, Second 2013.

[3] G. Piro, L. Grieco, G. Boggia, F. Capozzi, and P. Camarda, "Simulating LTE cellular systems: an open-source framework," *IEEE Trans. Veh. Technol.*, vol. 60, no. 2, pp. 498 –513, Feb. 2011.

[4] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2014-2019," Tech. Rep., 2015.

[5] J. Erman and K. K. Ramakrishnan, "Understanding the super-sized traffic of the super bowl," in *Proceedings of the 2013 conference on Internet measurement conference*. ACM, 2013, pp. 353–360.

[6] M. M. Belleville, E. E. Cantatore, H. H. Fanet, P. P. Fiorini, P. P. Nicole, M. Pelgrom, C. C. Piguet, R. R. Hahn, C. C. Van Hoof, R. Vullers *et al.*, "Energy autonomous systems: future trends in devices, technology, and systems," 2009.

[7] Auri Aittokallio, "Nokia unveils radio cloud architecture." [Online]. Available: `http://telecoms.com/399321/nokia-unveils-radio-cloud-architecture/`

[8] M. Kalil, A. Shami, and A. Al-Dweik, "Qos-aware power-efficient scheduler for LTE uplink," *IEEE Trans. Mobile Comput.*, vol. PP, no. 99, pp. 1–1, Oct. 2014.

[9] S. Sesia, I. Toufik, and M. Baker, *LTE - The UMTS Long Term Evolution: From Theory to Practice*. Wiley Publishing, 2009.

[10] G. Miao, N. Himayat, G. Li, and S. Talwar, "Low-complexity energy-efficient scheduling for uplink OFDMA," *IEEE Trans. Commun.*, vol. 60, no. 1, pp. 112–120, Jan. 2012.

[11] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks-part II: algorithm development," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 625–634, Mar. 2005.

[12] A. Marques, L. Lopez-Ramos, G. Giannakis, J. Ramos, and A. Caama Ando, "Optimal cross-layer resource allocation in cellular networks using channel- and queue-state information," *IEEE Trans. Veh. Technol.*, vol. 61, no. 6, pp. 2789 –2807, July 2012.

[13] A. Ahmad and M. Assaad, "Polynomial-complexity optimal resource allocation framework for uplink SC-FDMA systems," in *IEEE Global Commun. Conf. (GLOBECOM)*, pp. 1-5, Dec. 2011.

[14] K. Yang, N. Prasad, and X. Wang, "A Message-Passing Approach to Distributed Resource Allocation in Uplink DFT-Spread-OFDMA Systems," *IEEE Trans. Commun.*, vol. 59, no. 4, pp. 1099–1113, Apr. 2011.

[15] N. Prasad, H. Zhang, H. Zhu, and S. Rangarajan, "Multiuser scheduling in the 3GPP LTE cellular uplink," *IEEE Trans. Mobile Comput.*, vol. 13, no. 1, pp. 130–145, Jan. 2014.

[16] M. Kalil, J. Samarabandu, A. Shami, and A. Al-Dweik, "Performance evaluation of genetic algorithms for resource scheduling in LTE uplink," in *Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Aug 2014, pp. 948–952.

[17] F. Ren, Y. Xu, H. Yang, J. Zhang, and C. Lin, "Frequency domain packet scheduling with stability analysis for 3GPP LTE uplink," *IEEE Trans. Mobile Comput.*, vol. 12, no. 12, pp. 2412–2426, Dec. 2013.

[18] E. Prabhakar, B. Biyikoglu and A. El Gamal, "Energy-efficient transmission over a wireless link via lazy packet scheduling," in *Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2001, pp. 386–394.

[19] S. Lee, I. Pefkianakis, A. Meyerson, S. Xu, and S. Lu, "Proportional fair frequency-domain packet scheduling for 3GPP LTE uplink," in *IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2009, pp. 2611–2615.

[20] D. Dechene and A. Shami, "Energy-aware resource allocation strategies for LTE uplink with synchronous HARQ constraints," *IEEE Trans. Mobile Comput.*, vol. 13, no. 2, pp. 422–433, Feb. 2014.

[21] D. Dechene and A. Shami, "Energy efficient resource allocation in SC-FDMA uplink with synchronous HARQ constraints," in *IEEE International Conference on Communications (ICC)*, June 2011, pp. 1–5.

[22] ——, "Energy efficient QoS constrained scheduler for SC-FDMA uplink," *Physical Communication*, vol. 8, no. 0, pp. 81 – 90, Sept. 2013.

[23] A. Aijaz, X. Chu, and A. Aghvami, "Energy efficient design of SC-FDMA based uplink under QoS constraints," *IEEE Wireless Commun. lett.*, vol. PP, no. 99, pp. 1–4, Jan. 2014.

[24] A. T. Harri Holma, *LTE for UMTS: Evolution to LTE-Advanced*. John Wiley & Sons, 2011.

[25] M. Kalil, A. Shami, and A. Al-Dweik, "Power-efficient QoS scheduler for LTE uplink," in *IEEE Int. Conf. Commun. (ICC)*, June 2013, pp. 6200–6204.

[26] G. Song, Y. Li, and L. Cimini, "Joint channel-and queue-aware scheduling for multiuser diversity in wireless OFDMA networks," *IEEE Trans. Commun.*, vol. 57, no. 7, pp. 2109–2121, July 2009.

[27] M. Neely and S. Supittayapornpong, "Dynamic Markov decision policies for delay constrained wireless scheduling," *IEEE Trans. Autom. Control*, vol. 58, no. 8, pp. 1948–1961, Aug. 2013.

[28] M. Zafer and E. Modiano, "Minimum energy transmission over a wireless channel with deadline and power constraints," *IEEE Trans. Autom. Control*, vol. 54, no. 12, pp. 2841–2852, Dec. 2009.

[29] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[30] M. Zafer and E. Modiano, "A calculus approach to energy-efficient data transmission with Quality-of-Service constraints," *IEEE/ACM Trans. Networking*, vol. 17, no. 3, pp. 898–911, June 2009.

[31] H. Ekstrom, "QoS control in the 3GPP evolved packet system," *IEEE Commun. Mag.*, vol. 47, Feb. 2009.

[32] S. Shakkottai and A. L. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in HDR," *Bell labs technical reports*, 2001.

[33] M. M. Nasralla and M. G. Martini, "A downlink scheduling approach for balancing QoS in LTE wireless networks," in *Personal Indoor and Mobile Radio Commun. (PIMRC)*, Sept. 2013, pp. 1571–1575.

[34] M. Kalil, A. Shami, A. Al-Dweik, and S. Muhaidat, "Low-complexity power-efficient schedulers for lte uplink with delay-sensitive traffic," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4551–4564, Oct 2015.

[35] L. Zhang, B. Tiwana, Z. Qian, Z. Wang, R. P. Dick, Z. M. Mao, and L. Yang, "Accurate online power estimation and automatic battery behavior based power model generation for smartphones," in *Proc. 8th IEEE/ACM/IFIP int. conf. Hardware/softw. codesign and syst. synthesis*, 2010, pp. 105–114.

[36] N. Ding, D. Wagner, X. Chen, Y. C. Hu, and A. Rice, "Characterizing and modeling the impact of wireless signal strength on smartphone battery drain," in *Proc. of the*

ACM SIGMETRICS/int. conf. Measurement and modeling of comput. syst. ACM, 2013, pp. 29–40.

[37] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "A close examination of performance and power characteristics of 4G LTE networks," in *Proc. 10th int. conf. Mobile syst. applications services*, 2012, pp. 225–238.

[38] E. Yaacoub and Z. Dawy, "A survey on uplink resource allocation in OFDMA wireless networks," *IEEE, Commun. Surveys Tutorials*, vol. 14, no. 2, pp. 322–337, May 2012.

[39] T. Wang, F. Glineur, J. Louveaux, and L. Vandendorpe, "Weighted sum rate maximization for downlink OFDMA with subcarrier-pair based opportunistic DF relaying," *IEEE Trans. Signal Processing*, vol. 61, no. 10, pp. 2512–2524, May 2013.

[40] C. Y. Wong, R. Cheng, K. Lataief, and R. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation ," *IEEE JSAC*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.

[41] H. Boche and M. Schubert, "Nash bargaining and proportional fairness for wireless systems," *IEEE/ACM Trans. Networking*, vol. 17, no. 5, pp. 1453–1466, Oct. 2009.

[42] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans Commun.*, vol. 54, no. 7, pp. 1310–1322, July 2006.

[43] P. Liu, R. Berry, and M. Honig, "A fluid analysis of a utility-based wireless scheduling policy," *IEEE Trans. Inform. Theory*, vol. 52, no. 7, pp. 2872–2889, July 2006.

[44] M. Shaat and F. Bader, "Asymptotically optimal resource allocation in OFDM-based cognitive networks with multiple relays," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 892–897, Mar. 2012.

[45] V. Lau and C. H. Koh, "Tradeoff analysis of delay-power-CSIT quality of dynamic backpressure algorithm for energy efficient OFDM systems," *IEEE Trans. Signal Processing*, vol. 60, no. 8, pp. 4254–4263, Aug. 2012.

[46] M. Mehrjoo, S. Moazeni, and X. S. Shen, "Resource allocation in OFDMA networks based on interior point methods," *Wireless Commun. and Mobile Comput.*, vol. 10, no. 11, pp. 1493–1508, Nov. 2010.

[47] I. Wong, O. Oteri, and W. Mccoy, "Optimal resource allocation in uplink SC-FDMA systems," *IEEE Trans. on Wireless Commun.*, vol. 8, no. 5, pp. 2161 –2165, May 2009.

[48] O. Nwamadi, X. Zhu, and A. Nandi, "Dynamic physical resource block allocation algorithms for uplink long term evolution," *IET Commun.*, vol. 5, no. 7, pp. 1020–1027, May 2011.

[49] J. Kim, D. Kim, and Y. Han, "Proportional fair scheduling algorithm for SC-FDMA in LTE uplink," in *IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2012, pp. 4816–4820.

[50] M. Al-Rawi, R. Jantti, J. Torsner, and M. Sagfors, "On the performance of Heuristic opportunistic scheduling in the uplink of 3G LTE networks," in *IEEE Int. Symp. Personal Indoor Mobile Radio Commun. (PIMRC)*, 2008.

[51] X. Xiang, C. Lin, X. Chen, and X. Shen, "Toward optimal admission control and resource allocation for LTE-A femtocell uplink," *IEEE Trans. Veh. Technol.*, vol. PP, no. 99, pp. 1–1, 2014.

[52] H. Tabassum, Z. Dawy, M. Alouini, and F. Yilmaz, "A generic interference model for uplink OFDMA networks with Fractional Frequency Reuse," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1491–1497, March 2014.

[53] 3GPP R1-050718, "Simulation methodology for EUTRA UL: IFDMA and DFT-spread-OFDMA," 2005.

[54] C. Xiong, G. Li, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient resource allocation in OFDMA networks," *IEEE Trans. Commun*, vol. 60, no. 12, pp. 3767–3778, Dec. 2012.

[55] R. Xie, F. Yu, H. Ji, and Y. Li, "Energy-efficient resource allocation for Heterogeneous cognitive radio networks with Femtocells," *IEEE Trans. Wireless Commun*, vol. 11, no. 11, pp. 3910–3920, November 2012.

[56] 3GPP TS 23.203 V9.9.0, "Policy and charging control architecture (Release 9)," 2011.

[57] A. Khrwat, B. S. Sharif, C. Tsimenidis, S. Boussakta, and A. J. Al-Dweik, "Channel prediction for limited feedback precoded MIMO-OFDM systems," in *IEEE Int. Symp. sign. Proces. and Inform. Technol. (ISSPIT)*, Dec. 2009, pp. 195–200.

[58] A. Duel-Hallen, "Fading channel prediction for mobile radio adaptive transmission systems," *Proc. IEEE*, vol. 95, no. 12, pp. 2299–2313, Dec. 2007.

[59] 3GPP TS 36.331 v 11.8.0, "Universal Terrestrial Radio Access (UTRA); Radio Resource Control (RRC); Protocol specification," 2011.

[60] O. Sundstrom and L. Guzzella, "A generic dynamic programming Matlab function," in *IEEE Control Applications, (CCA) Intelligent Control, (ISIC)*, July 2009, pp. 1625–1630.

[61] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, "Cellslice: Cellular wireless resource slicing for active RAN sharing," in *5th International Conference on Communication Systems and Networks (COMSNETS'13)*, Jan. 2013, pp. 1–10.

[62] M. Kalil, A. Shami, and Y. Yinghua, "Wireless resources virtualization in LTE systems," in *IEEE Int. Conf. Comput. Commun. (INFOCOM), Workshop Mobile Cloud Comput.*, May 2014.

[63] C. Han, T. Harrold, S. Armour, I. Krikidis, S. Videv, P. M. Grant, H. Haas, J. Thompson, I. Ku, C.-X. Wang, T. A. Le, M. Nakhai, J. Zhang, and L. Hanzo, "Green radio: radio techniques to enable energy-efficient wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 46–54, June 2011.

[64] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC)," *IEEE Network,*, vol. 28, no. 6, pp. 18–26, Nov 2014.

[65] R. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems," *Communications Magazine, IEEE*, vol. 52, no. 5, pp. 94–101, May 2014.

[66] Orange, "Orange fast facts - networks," 2013.

[67] Y. Zaki, L. Zhao, C. Goerg, and A. Timm-Giel, "LTE wireless virtualization and spectrum management," in *Wireless and Mobile Netw. Conf. (WMNC)*, Oct. 2010.

[68] M. Li, L. Zhao, X. Li, X. Li, Y. Zaki, A. Timm-Giel, and C. Gorg, "Investigation of network virtualization and load balancing techniques in LTE networks," in *IEEE Veh. Technol. Con. (VTC)*, May 2012.

[69] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, "NVS: A substrate for virtualizing wireless resources in cellular networks," *IEEE/ACM Trans. Netw*, vol. 20, no. 5, pp. 1333–1346, Oct 2012.

[70] B. Leonardo, F. Torsten, G. Francesco, K. Eleftherios, L. Remcom, A. Osman, P. Irene, S. Michal, and Z. Haibin, "Resource sharing among wireless network operators: Spectrum, infrastructure, and full radio access network sharing," *SAPHYRE White Paper #3*, 2013.

[71] M. Rahman and H. Yanikomeroglu, "Enhancing cell-edge performance: a downlink dynamic interference avoidance scheme with inter-cell coordination," *IEEE Trans. Wireless Commun.*, vol. 9, no. 4, pp. 1414–1425, Apr. 2010.

[72] Y. Ma and D. I. Kim, "Rate-maximization scheduling schemes for uplink OFDMA," *IEEE Trans. on Wireless Commun.*, vol. 8, no. 6, pp. 3193–3205, Jun. 2009.

[73] U. Paul, A. Subramanian, M. Buddhikot, and S. Das, "Understanding traffic dynamics in cellular data networks," in *IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2011, pp. 882–890.

[74] R. Mahindra, M. Khojastepour, H. Zhang, and S. Rangarajan, "Radio access network sharing in cellular networks," in *IEEE Int. Conf. Netw. Protocols (ICNP)*, Oct. 2013, pp. 1–10.

[75] "C-RAN: the road towards green radio access network. presentation," China Mobile Research Institute, Arlington, VA, USA, Tech. Rep., 2011. [Online]. Available: `http://http://labs.chinamobile.com/cran/wp-content/uploads/2014/06/`

[76] A. Checko, H. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. Berger, and L. Dittmann, "Cloud RAN for mobile networks-a technology overview," *IEEE Commun. Surveys Tuts.*, vol. PP, no. 99, pp. 1–1, Sept. 2014.

[77] S. Namba, T. Matsunaka, T. Warabino, S. Kaneko, and Y. Kishi, "Colony-RAN architecture for future cellular network," in *Future Network Mobile Summit (FutureNetw), 2012*, July 2012, pp. 1–8.

[78] M. Madhavan, P. Gupta, and M. Chetlur, "Quantifying multiplexing gains in a wireless network cloud," in *IEEE Int. Commun. Conf. (ICC)*, June 2012, pp. 3212–3216.

[79] "Active RAN sharing could save $60 billion for operators." [Online]. Available: `http://www.cellular-news.com/story/36831.php`

[80] C. Liang and F. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 358–380, Firstquarter 2015.

[81] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks-part I: theoretical framework," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 614–624, Mar. 2005.

[82] L. Zhao, M. Li, Y. Zaki, A. Timm-Giel, and C. Gorg, "LTE virtualization: from theoretical gain to practical solution," in *Int. Teletraffic Congr. (ITC)*, Sept. 2011.

[83] F. Fu and U. Kozat, "Stochastic game for wireless network virtualization," *IEEE/ACM Trans. Netw*, vol. 21, no. 1, pp. 84–97, Feb. 2013.

[84] V. Jumba, S. Parsaeefard, M. Derakhshani, and T. Le-Ngoc, "Resource provisioning in wireless virtualized networks via Massive-MIMO," *IEEE Wireless Commun. Lett.*, vol. 4, no. 3, pp. 237–240, June 2015.

[85] A. Hamza, S. Khalifa, H. Hamza, and K. Elsayed, "A survey on inter-cell interference coordination techniques in OFDMA-based cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 1642–1670, Nov. 2013.

[86] 3GPP TR 36.921 , "Home eNode B (HeNB) Radio Frequency (RF) requirements analysis," 2011.

[87] G. Boudreau, J. Panicker, N. Guo, R. Chang, N. Wang, and S. Vrzic, "Interference coordination and cancellation for 4G networks," *IEEE Commun. Magazine*, vol. 47, no. 4, pp. 74–81, Apr. 2009.

[88] T. Novlan, R. Ganti, A. Ghosh, and J. Andrews, "Analytical evaluation of fractional frequency reuse for OFDMA cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 4294–4305, Dec. 2011.

[89] K. Wang, M. Zhao, and W. Zhou, "Graph-based dynamic frequency reuse in Cloud-RAN," in *IEEE Wireless Commun. and Netw. Conf. (WCNC)*, Apr. 2014, pp. 105–110.

[90] Y. Shi, J. Zhang, and K. Letaief, "Group sparse beamforming for green Cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, May 2014.

[91] A. Davydov, G. Morozov, I. Bolotin, and A. Papathanassiou, "Evaluation of joint transmission CoMP in C-RAN based LTE-A HetNets with large coordination areas," in *IEEE Globecom Workshops*, Dec. 2013, pp. 801–806.

[92] M. Lau, W. Yue, and L. Ping, "Equal power allocation of IDMA systems: Feasibility, optimality, and throughput," *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, pp. 746–756, Feb 2009.

[93] B. Luo, Q. Cui, H. Wang, and X. Tao, "Optimal joint water-filling for coordinated transmission over frequency-selective fading channels," *IEEE Commun. Lett.*, vol. 15, no. 2, pp. 190–192, February 2011.

[94] 3GPP TR 25.814 V7.1.0, "Physical layer aspects forevolved Universal Terrestrial Radio Access (UTRA) (Release 7) ," 2007.

[95] V. Chandrasekhar and J. Andrews, "Spectrum allocation in tiered cellular networks," *IEEE Trans. Wireless Commun.*, vol. 57, no. 10, pp. 3059–3068, Oct. 2009.

[96] A. Leith, M.-S. Alouini, D. I. Kim, X. Shen, and Z. Wu, "Flexible proportional-rate scheduling for OFDMA system," *IEEE Trans. Mobile Comput.*, vol. 12, no. 10, pp. 1907–1919, Oct. 2013.

[97] M. Katoozian, K. Navaie, and H. Yanikomeroglu, "Utility-based adaptive radio resource allocation in OFDM wireless networks with traffic prioritization," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 66–71, Jan 2009.

[98] G. Lim, C. Xiong, L. J. Cimini, and G. Y. Li, "Energy-efficient resource allocation for OFDMA-Based Multi-RAT networks," *Wireless Communications, IEEE Transactions on*, vol. 13, no. 5, pp. 2696–2705, May 2014.

Curriculum Vitae

Name: Mohamad Kalil

Post-secondary Education and Degrees:

2012-2015 Ph.D
Electrical and Computer Engineering
Western University
London, Ontario, Canada

2004-2009 B.Sc., 2009-2011 M.Sc.
Electrical Engineering
Jordan University of Science and Technology
Irbid, Jordan

Related Work Experience

2012-2015
Teaching Assistance
Western University
London, Ontario, Canada

ECE 4436 - Networking: Principles, Protocols, and Architectures
SE 3314 - Computer Networks and Applications
ECE 2241 - Electrical Laboratory II

Honours and Awards

Graduate Student Award for Excellence in Research
The Department of Electrical and Computer Engineering
Western University, London, Ontario, Canada, 2015

Best Paper Award
International Wireless Communications and Mobile Computing Conference (IWCMC), 2014

Graduate Fellowship
The Department of Electrical and Computer Engineering
Western University, London, Ontario, Canada, 2012-2015

Publications

[J1] M. Kalil, A. Shami, and A. Al-Dweik, ”QoS-Aware Power-Efficient Scheduler for LTE Uplink”, IEEE Transactions on Mobile Computing, 2014

[J2] M. Kalil, A. Shami, A. Al-Dweik, and S. Muhaidat, ”Low-Complexity Power-Efficient Schedulers for LTE Uplink with Delay-Sensitive Traffic”, IEEE Transactions on Vehicular Technology, 2015

[J3] M. Kalil, A. Moubayed, A. Shami, and A. Al-Dweik, ” Optimal Low-Complexity Scheduler for Wireless Resource Virtualization”, IEEE Wireless Communications Letters (in press)

[C1] M. Kalil, A. Shami, and A. Al-Dweik, ”Power-efficient QoS scheduler for LTE uplink”, IEEE ICC, 2013

[C2] M. Kalil, A. Shami, and Y. Yinghua, ”Wireless resources virtualization in LTE systems”, in IEEE INFOCOM Workshop on Mobile Cloud Computing, 2014

[C3] M. Kalil, J. Samarabandu, A. Shami, A. Al-Dweik, ”Performance Evaluation of Genetic Algorithms for Resource Scheduling in LTE Uplink”, IWCMC, 2014 (Received Best Paper Award)

[C4] M. Kalil, Mohammad M. Banat, and Faouzi Bader, ”Three Dimensional Pilot Aided Channel Estimation for Filter Bank Multicarrier MIMO Systems with Spatial Channel Correlation”, JIEEEC2013, April 2013. Amman, Jordan

[B1] M. Kalil , K. A. Meerja, A. Refaey, and A. Shami, ”Virtual Mobile Networks in Clouds” in book ”Advances in Mobile Cloud Computing Systems” to be published in 2015 by CRC Press (Taylor & Francis Group)